# Conversational Crowdsourcing

Qiu, S.

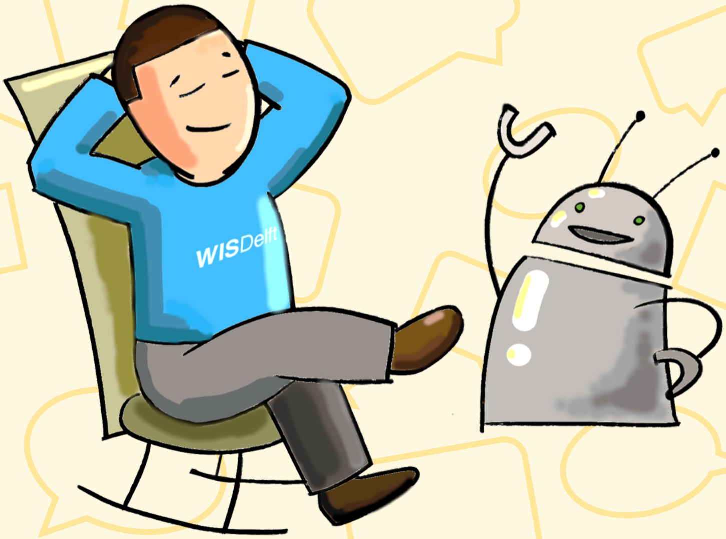
**DOI**
10.4233/uuid:d37db2c0-cf16-4edf-97ba-aebff35011b5

**Publication date**
2021

**Document Version**
Final published version

**Citation (APA)**
Qiu, S. (2021). *Conversational Crowdsourcing*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:d37db2c0-cf16-4edf-97ba-aebff35011b5

# CONVERSATIONAL CROWDSOURCING



SIHANG QIU

# Conversational Crowdsourcing

Sihang Qiu

# Conversational Crowdsourcing

## Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus prof.dr.ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Monday 4 October 2021 at 10:00 o'clock

by

**Sihang QIU**

Master of Control Science and Engineering
National University of Defense Technology, China
born in Zhejiang, China.

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

| | |
|---|---|
| Rector Magnificus, | chairperson |
| Prof.dr.ir. G.J.P.M. Houben | Delft University of Technology, promotor |
| Prof.dr.ir. A. Bozzon | Delft University of Technology, promotor |
| Dr. U.K. Gadiraju | Delft University of Technology, copromotor |
| | |
| *Independent members:* | |
| Prof.dr. P.S. César Garcia | Delft University of Technology and<br>Centrum Wiskunde & Informatica |
| Prof.dr.ir. S.P. Hoogendoorn | Delft University of Technology |
| Prof.dr. P. Markopoulos | Eindhoven University of Technology |
| Dr. S. Hosio | University of Oulu |
| Prof.dr. G. Kortuem | Delft University of Technology, reserve member |

SIKS Dissertation Series No. 2021-22
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Published and distributed by: Sihang Qiu
E-mail: sihangq@acm.org

*Keywords:* Crowdsourcing, Conversational agent, Chatbot, Engagement, Satisfaction, Mood, Self-identification, Avatar, Memorability, Health

*Cover design by:* Sihang Qiu

*Printed and bound in The Netherlands by:* ProefschriftMaken

ISBN: 978-94-6423-487-9

# Summary

Crowdsourcing has become a standard approach for the collection of the human input required by scientists and practitioners alike to execute their experiments, or to train, control, and verify the behavior of their intelligent systems. Despite years of successful research and industrial application, how to improve the engagement and satisfaction of crowd workers with crowdsourcing tasks is still an open research question. In this thesis, we introduce conversational crowdsourcing – a novel crowdsourcing interaction paradigm based on conversational interfaces. We study conversational crowdsourcing, and experimentally evaluate its ability to foster workers' engagement and satisfaction from four perspectives: conversational crowdsourcing design, improving worker engagement and satisfaction, analyzing the roles of worker mood and self-identification, and applying conversational crowdsourcing for conducting online studies.

We describe the design of conversational crowdsourcing and show that conversational crowdsourcing can achieve similar output quality and execution time compared to the traditional web-based crowdsourcing. To facilitate our research, we designed and developed TickTalkTurk, a web application that facilitates the design and development of conversational crowdsourcing tasks on popular crowdsourcing platforms.

We demonstrate the feasibility of improving worker engagement and satisfaction and show that conversational crowdsourcing can improve worker retention and perceived engagement that are significantly connected to satisfaction. We present a reliable conversational style estimation method and illustrate that style estimation can be a useful tool for facilitating outcome prediction and task assignment.

Since conversation is strongly associated with human emotions, we investigate the roles of emotional factors, including worker mood and self-identification. We show that conversational crowdsourcing can improve worker retention irrespective of worker moods, and different conversational styles exhibit the potential to improve engagement of workers in different moods. Our study reveals the occurrence of similarity identification and wishful identification in conversational crowdsourcing and the feasibility of using worker avatars with conversational crowdsourcing for reducing cognitive workload.

We show the suitability of conversational crowdsourcing as a research tool in the context of information retrieval and workers' health research. Our findings reveal that conversational interfaces have the potential to help users better retain information consumed. We also apply conversational crowdsourcing to carry out a survey study to understand worker health on popular crowdsourcing platforms. We show that worker health is related to both physical and psychosocial working environments.

With our work, we show that conversational crowdsourcing has the potential to create a better working environment for workers operating on online microtask crowdsourcing platforms. The thesis concludes with a discussion on the implications of our work, and with the identification of several directions for further investigation.

# Samenvatting

Crowdsourcing is een standaardaanpak geworden voor de verzameling van menselijke input die wetenschap en praktijk nodig hebben voor het uitvoeren van hun experimenten of om hun intelligente systemen te trainen, te controleren of het gedrag te verifiëren. Ondanks jaren van succesvolle toepassing in onderzoek en praktijk is het verbeteren van engagement en tevredenheid van crowd-werkers met crowdsourcing-taken nog steeds een open onderzoeksvraag. In dit proefschrift presenteren wij conversational crowdsourcing – een nieuw interactieparadigma voor crowdsourcing dat gebaseerd is op conversatie-interfaces. Wij onderzoeken conversational crowdsourcing en evalueren met behulp van experimenten in hoeverre het engagement en tevredenheid van werkers kan bevorderen vanuit vier verschillende perspectieven: ontwerpen van conversational crowdsourcing, verbeteren van engagement en tevredenheid van werkers, analyseren van de rollen van de stemming van werkers en zelfidentificatie, en toepassen van conversational crowdsourcing voor het uitvoeren van online-studies.

Wij beschrijven het ontwerp van conversational crowdsourcing en laten zien dat conversational crowdsourcing een effectief alternatief ten opzichte van traditionele web-gebaseerde crowdsourcing kan zijn in termen van outputkwaliteit en executietijd. Om ons onderzoek te vergemakkelijken hebben wij TickTalkTurk ontworpen en ontwikkeld, een webapplicatie die het ontwerp en de ontwikkeling van conversational crowdsourcing-taken op populaire crowdsourcing-platforms vergemakkelijkt.

Wij tonen de haalbaarheid van het verbeteren van engagement en tevredenheid van werkers en laten zien dat conversational crowdsourcing retentie en waargenomen engagement van werkers kan verbeteren, wat significant gerelateerd is aan tevredenheid. Wij presenteren een betrouwbare methode voor de schatting van conversatiestijl en lichten toe dat stijlschatting een nuttig middel kan zijn om de uitkomstvoorspelling en taakverdeling te ondersteunen.

Omdat conversatie sterk gerelateerd is aan menselijke emoties, onderzoeken wij de rollen van emotionele factoren zoals de stemming werkers en zelfidentificatie. Wij laten zien dat conversational crowdsourcing de retentie van werkers onafhankelijk van hun stemming kan verbeteren en dat verschillende conversatiestijlen de mogelijkheid bieden voor potentiële verbetering van engagement van werkers in verschillende stemmingen. Ons onderzoek toont hoe overeenkomstidentificatie en wenselijke identificatie in conversational crowdsourcing voorkomen en toont de haalbaarheid van het gebruik van avatars met conversational crowdsourcing voor het verminderen van cognitieve werklast.

Wij laten de geschiktheid van conversational crowdsourcing zien als onderzoeksgereedschap in de context van information retrieval en onderzoek over de gezondheid van werkers. Onze bevindingen tonen dat conversatie-interfaces het potentieel hebben om gebruikers te helpen om geconsumeerde informatie beter te onthouden. Daarnaast hebben wij conversational crowdsourcing ook toegepast op een survey-studie over het begrijpen van de gezondheid van werkers. Wij laten zien dat gezondheid van werkers gerelateerd is aan zowel fysieke als psychosociale werkomgevingen.

Door middel van dit werk laten wij zien dat conversational crowdsourcing het potentieel heeft om een betere werkomgeving te creëren voor workers op online microtaak crowdsourcing-platforms. Het proefschrift sluit af met een discussie over de implicaties van ons werk en met de identificatie van richtingen voor toekomstig onderzoek.

# Acknowledgements

I would like to deliver my gratitude to all who have given me guidance, help, and joy, to overcome the challenges that I encountered in my journey of pursuing a PhD.

First of all, I would like to express my highest gratitude to my promotors Geert-Jan Houben and Alessandro Bozzon. Thank you, Geert-Jan, for giving me an opportunity to start my PhD in the Netherlands, and for providing me with your strong support. Thank you, Alessandro, for your extremely helpful and useful advice, making my four-year PhD journey delightful and enjoyable. I also want to thank my copromotor Ujwal Gadiraju. It was a great pleasure working with you. I cherish the memory of fun discussions and online gatherings we had together for each of our publications.

I would like to thank Prof. Pablo César, Prof. Serge Hoogendoorn, Prof. Panos Markopoulos, Dr. Simo Hosio, and Prof. Gerd Kortuem, for accepting to be my committee members and providing me with valuable feedback.

I am particularly grateful to Jie Yang, Oana Inel, Shahin Sharifi, Ioannis Petros Samiotis, Agathe Balayn, Gaole He, Sepideh Mesbah, Andrea Mauri, Vincent Gong, Achilleas Psyllidis, and Carlo van der Valk — my colleagues currently or formerly working in Sigma/Kappa team. Many thanks to the members of the Web Information Systems (WIS): Marcus Specht, Nava Tintarev, Claudia Hauff, Asterios Katsifodimos, Christoph Lofi, Yue Zhao, Guanliang Chen, Dan Davis, Emily Sullivan, Mesut Kaya, Panagiotis Mavridis, David Maxwell, Marios Fragkoulis, Rihan Hai, Lixia Chu, Alisa Rieger, Andra Ionescu, Arthur Câmara, Christos Koutras, Felipe Moraes, Georgios Siachamis, Gustavo Penha, Huiyuan Lai, Kyriakos Psarakis, Manuel Valle Torre, Nirmal Roy, Peide Zhu, Sara Salimzadeh, Shabnam Najafian, Tim Draws, Ziyu Li, and Daphne Stephan.

I would like to thank my amazing office mates: Yue, Sepideh, Shahin, Shabnam, Petros, Agathe, Christos, Andrea, and Manuel, for the happy moments we shared. I would like to thank Jie and Guanliang for numerous valuable suggestions. I want to thank Tim for translating my thesis summary into Dutch. I would like to express my sincere gratitude to Daphne for helping me throughout the final stage of my PhD.

I would also like to thank a number of talented students: Owen Huang, Enreina Annisa Rizkiasri, Neha Thuraka, Gerard van Alphen, Shipra Sharma, Orestis Kanaris, Bar Lerer, Emilija Zlatkutė, Just van der Veeken, Jesse Jansen, and Ji-Youn Jung.

I really appreciate spending time with my Chinese friends in the Netherlands. Special thanks must go to Xiaohui, Hai, Qi, Hanqing, Baozhou, Hè, and Guishan.

I would like to thank my supervisors and mentors when I was studying in China: Prof. Xiaogang Qiu, Dr. Bin Chen, and Prof. Yingmei Wei. Thank you for leading me to start scientific research.

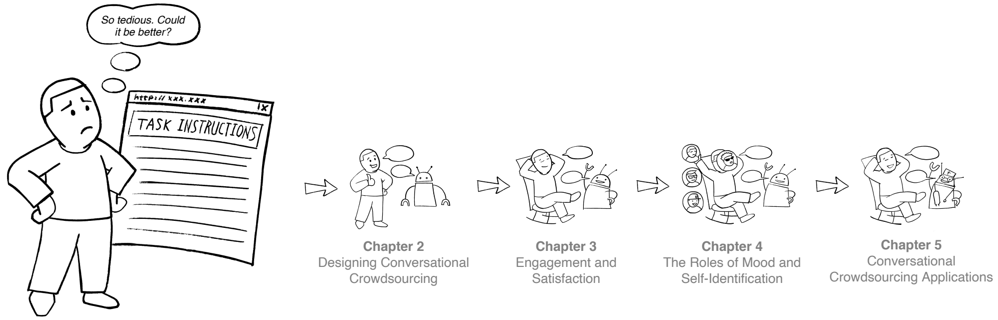
Finally, I would like to express my deepest gratitude to my parents for their never-ending love, encouragement, and support.

Sihang Qiu
September, 2021
Rotterdam, the Netherlands

# Contents

# Chapter 1

# Introduction


So tedious. Could it be better?
http://xxx.xxx
TASK INSTRUCTIONS
Chapter 2
Designing Conversational Crowdsourcing
Chapter 3
Engagement and Satisfaction
Chapter 4
The Roles of Mood and Self-Identification
Chapter 5
Conversational Crowdsourcing Applications

The role of human input is widely acknowledged to be essential for the design, development, and control of systems that include artificial intelligence components [46, 83], for instance, for training datasets creation, systems evaluation, computer supported cooperative work, and experimentation. Crowdsourcing has become a primary means to effectively collect human input from anonymous users of the Internet, leading to the prosperity of crowdsourcing marketplaces (such as Amazon's Mechanical Turk[1], Prolific[2], Toloka[3], and Appen[4]) that attract an increasing number of people, often working full-time. On a crowdsourcing platform, crowd workers can select and complete tasks offered by requesters who demand the data, to earn their monetary reward. Considering the great potential of crowdsourcing marketplaces, the leading scientists in the crowdsourcing domain have identified that the future of crowdsourcing will depend on both worker satisfaction (motivation, feedback, pay, etc.) and organizational performance (job design, task decomposition, career ladder, etc.) [117]. Previous work has extensively focused on issues of worker performance modeling [19, 237, 93, 243, 123] and quality control [43, 71, 127]. However, worker satisfaction and engagement received comparatively less attention.

Traditionally, crowdsourcing tasks are firstly designed by requesters, and then executed by crowd workers using web-based interfaces. Crowdsourcing tasks are generally designed in a way that makes workers perform tasks in long and repetitive batches [51]. Recent studies

[1] https://www.mturk.com/

[2] https://www.prolific.co/

[3] https://toloka.ai/

[4] https://appen.com

have revealed that crowdsourcing in such a monotonous way can lead to effects such as boredom, fatigue, and high drop-out rates [87, 148]. Such effects can negatively affect worker satisfaction and engagement, and imperceptibly set invisible barriers for participation.

To further lower the barrier for participants and improve worker satisfaction, prior studies have tried a variety of task- and context-specific means [139, 55, 182]. Although standalone solutions exist for specific task types, we still look forward to solutions for better engaging workers that are easily applicable for a variety of task types. Similar issues of satisfaction and engagement have been observed in other areas related to human-computer interactions, such as online learning [239]. To address such issues, conversational interfaces emerged as a powerful approach aiming to provide seamless means of interaction with virtual assistants, chatbots, or messaging services. Messaging applications using conversational interfaces have been reported to be more popular than conventional social networks [201], resulting in a

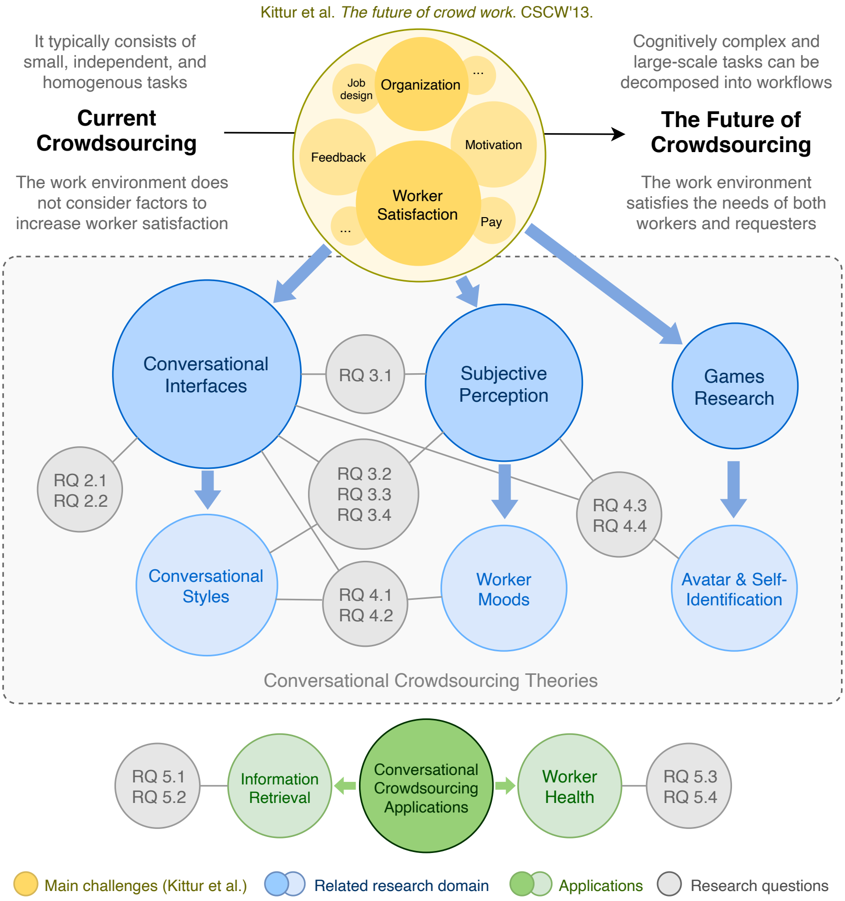

Kittur et al. The future of crowd work. CSCW'13.
It typically consists of small, independent, and homogenous tasks
Current Crowdsourcing
The work environment does not consider factors to increase worker satisfaction
Job design
Organization
...
Feedback
Motivation
Worker Satisfaction
Pay
...
Cognitively complex and large-scale tasks can be decomposed into workflows
The Future of Crowdsourcing
The work environment satisfies the needs of both workers and requesters
Conversational Interfaces
RQ 3.1
Subjective Perception
Games Research
RQ 2.1
RQ 2.2
RQ 3.2
RQ 3.3
RQ 3.4
RQ 4.3
RQ 4.4
Conversational Styles
RQ 4.1
RQ 4.2
Worker Moods
Avatar & Self-Identification
Conversational Crowdsourcing Theories
RQ 5.1
RQ 5.2
Information Retrieval
Conversational Crowdsourcing Applications
Worker Health
RQ 5.3
RQ 5.4
Main challenges (Kittur et al.)
Related research domain
Applications
Research questions

Figure 1.1: The research scope of the thesis. Yellow bubbles represent main challenges identified by Kittur et al. [117]. Blue bubbles represent research background explained in Section 1.1 (the size of the bubble represents the relevance to this thesis). Green bubbles represent applications. Grey bubbles represent research questions explained in Section 1.2.

growing familiarity of people with conversational interfaces. In addition, conversational interfaces have been argued to have advantages over traditional graphical user interfaces, due to a more human-like means of interaction that conversational interfaces can provide and the potential engagement that conversation can stimulate [153].

In this thesis, we introduce a novel crowdsourcing interaction paradigm — conversational crowdsourcing — that we hypothesize to be able to improve the satisfaction and engagement of crowd workers on crowdsourcing marketplaces. In addition to the potential benefits for engagement and satisfaction mentioned above, user experience has been shown to be affected by conversational styles [113, 215] and human emotions in conversation [196]. We therefore study three important aspects related to 1) conversational styles, since prior works in linguistics have shown that conversational styles can play an important role in engaging speakers in inter-human communication [210, 211]; 2) worker moods, because related literature have revealed that emotions and moods can significantly affect user performance and engagement in either office work or online crowd work [217, 244]; and 3) self-identification, since it is strongly associated with emotions and it is effective in motivating users, inspired by games research [14]. Figure 1.1 presents the research scope of the thesis, where the bubbles in yellow represent the main challenges identified by Kittur et al. [117] with the aim of achieving a better future of crowd work; the bubbles in blue represent the related knowledge and research background; the bubbles in green represent the applications of conversational crowdsourcing; and the bubbles in grey represent the research questions being investigated in the thesis, and how they link with the related knowledge.

In the following sections, we will first introduce the research background from the aspects of lowering barriers for participation, conversational interfaces (including conversational styles), worker subjective perceptions (including worker moods), and games research (including self-identification with avatars) in Section 1.1. Then we will delve into several research questions to fill the knowledge gap (Section 1.2). In Section 1.3, we introduce the research methodologies used in this thesis.

## 1.1 Background

In this section, we elaborate related work from the perspectives of conversational interfaces, worker subjective perceptions, and gamification respectively, which have been shown to be related to worker satisfaction and engagement.

### Lowering Barriers for Participation in Microtask Crowdsourcing

Researchers proposed various methods to lower barriers for participation in crowdsourcing. Narula et al. designed *Mobileworks*, a mobile-based crowdsourcing platform that enable crowd workers to perform image recognition tasks [159]. The authors also show that in developing countries the crowd work marketplaces are often inaccessible. A prior work studied how crowdsourcing tasks could prevent workers (who have little digital literacy skills) from task execution and completion on Amazon's Mechanical Turk [112]. The authors found that the key usability barriers were the task instructions, user interface, and the cultural context. Therefore, the authors suggested that localization, simple user interfaces and task instructions should be considered in the task design, to better facilitate participation of crowd workers in India. Good task design and clear task instructions have been shown to be able to bring positive impacts to crowd work, and have been identified to be important factors to enable better access and participation from crowd workers [116, 140, 75]. Complementing these prior works, we propose to use conversational interfaces that people may be generally more familiar with as an alternative to standard web interfaces to lower participation barriers.

### Conversational Interfaces

Due to a more human-like interaction, conversational interfaces have been argued to have advantages over standard web-based user interfaces [153]. Therefore, we are currently witnessing a gradual rise of conversational interfaces are in various domains of our daily life [238]. Conversational interfaces can be served as a media for either inter-human conversation or human-machine conversation. As for human-machine conversation, a conversational interface usually features a conversational agent, which is a software programmed to automatically interpret and respond to requests expressed in natural language, so to mimic the behavior of a human interlocutor. *Chatbots* are a class of conversational agents that prevalently use text as a interaction medium. While research on chatbot systems dates back to the 1960s, the growing popularity of messaging platforms (especially on mobile devices) is sparking new interest both in industry and academia. In addition to the traditional focus on conversational purposes, recent work in Information Retrieval addressed informational task. For instance, Vtyurina et al. [226] investigate the use of a chatbot system as an alternative for search engines to retrieve information in a conversational manner. Avula et al. [5, 6] explored the adoption of chatbots for collaborative search and content recommendation. Vaccaro et al. [221] investigated the use of chatbot for styling personalization. Recent work has investigated the user experiences with regard to conversational interfaces, to understand user needs and satisfaction [35, 39, 136]. 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 [222]. These works have shown that conversational agents and interfaces can improve user experiences and have highlighted the need to further investigate the use of conversational agents in different scenarios.

***Conversational Interfaces for Crowdsourcing.*** Prior research has combined crowdsourcing and the conversational interface for training the dialogue manager or natural language processing component [122]. Lasecki et al. designed and developed Chorus, a conversational assistant able to assist users with general knowledge tasks [128]. 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 [99]. 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 [101]. Conversational microtask crowdsourcing is also deployed on social network platforms, combing with messaging applications, such as Facebook and Twitter. Savage et al. designed a platform named Botivist based on Twitter, engaging volunteers to action by using different strategies [193]. A previous work based on Facebook Messenger used a Chatbot to connect learners and experts, for providing experts' feedback to improve learners' work [218]. A conversational agent called Curious Cat was proposed to combine the crowdsourcing approach from a different perspective [20]. 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. Prior works have shown the utility of conversational systems in crowdsourcing. However, a novel conversational crowdsourcing paradigm which facilitates the execution of different types of popular crowdsourcing tasks is yet to be designed.

***Conversational Style and Work Outcomes.*** The conversational style is an essential element of a conversation, and it is also found to be related to work outcomes. The earliest systematic analysis of conversational style was performed by Lakoff [125]. She classified the stylistic strategies people used in everyday conversation into four categories, namely, *clarity*, *distance*, *deference*, and *camaraderie*. Lakoff found that speakers tend to use the camaraderie strategy when they want to be friendly, and use clarity strategy when they want to have the least relationship with another. Speakers can use a combination of different strategies in practice. Based on that, Tannen proposed a classification of conversational style that distributes speakers on a continuum from *High Consideratenes* to *High Involvement*. She also concluded important features and linguistic devices helping in classification of speakers' conversational styles [210, 211]. In terms of the usage of conversational style in human-computer interaction, Shamekhi et al. analyzed the preferred conversational style of users for a virtual agent [195]. They extrapolated the conversational style by indirectly asking users about their attitudes during a conversation. Thomas et al. analyzed styles of the information-seeking conversation from the MISC dataset [216] by using some measurable properties to represent features defined by Tannen, such as pronoun use, speech rate, pitch/loudness variation, and so on [215]. Conversational styles can also affect work outcomes and worker performance. Using Botivist [193], the authors analyzed how strategies (corresponding to different language styles) could potentially affect the outcome. Previous work evaluated the impact of linguistic style matching (LSM [81]) on team performance in long-term tasks [156]. Tausczik et al. designed a real-time language feedback system to test the work outcomes of student groups by monitoring communication patterns [213].

Many prior studies have used automated methods to predict and analyze age, gender and personality based on linguistic features [194, 2, 25, 161]. A recent study compared the impacts of conversational styles on online surveys [113]. Authors defined two styles, "casual" and "formal" respectively, and then applied these two styles on web platform and chatbot. They concluded that a chatbot could play a role as a human in surveys if an appropriate conversational style is used. Previous works about conversational styles however, are not directly applicable in microtask crowdsourcing.

### Subjective Perception in Crowdsourcing

The subjective perception of crowd workers plays important roles in microtask crowdsourcing with regard to many aspects. Previous works have studied subjective perceptions concerning worker enjoyment [37, 18], engagement [52, 139], and moods [244].

Worker engagement is crucial to microtask crowdsourcing since it has positive effects on building better relationships with crowd workers. Researchers have already noticed the importance of worker engagement and proposed methods to measure and predict it [139]. A previous work combined crowdsourcing with the process of learning [52], suggesting that both engagement and performance could be improved. The effort that workers make is also a major factor that can affect task execution time and cost. Cheng et al. proposed an effective way to measure worker efforts using "error-time area" (ETA), enabling a requester to rapidly evaluate the efficiency [33]. Apart from engagement and effort, the worker performance could be affected by more complex factors. Kazai et al. studied the relationship between workers' personality traits and crowdsourcing outcomes [109]. Considering the properties of outcomes such as accuracy and speed, workers can be classified into five main categories — Spammer, Sloppy, Incompetent, Competent and Diligent. Prior works also investigated the feasibility of using self-assessments to measure rather complex subjective properties like logical reasoning competence [70] and cognitive skill [91] — these subjective properties can significantly affect crowdsourcing results. Using such self-assessments before task execution could be useful for performance prediction and task assignment. However, we lack a thorough understanding of how workers' subjective perceptions such as satisfaction and experience related to tasks and their mental workload can be improved.

***Worker Moods in Crowdsourcing.*** Prior studies have shown that in real-life worker moods can affect people's task performance. In the context of crowd work, workers in a pleasant mood were also found to exhibit a better performance than those who were unpleasant [233, 241]. Others have shown that task execution time can also be affected by worker moods [151]. Recent work in the context of online crowdsourcing has revealed the relationship between worker moods and crowdsourcing task performance [244], where moods were measured using the Pick-A-Mood instrument [48]. Statistical tests indicated that worker moods had significant effects on their engagement. Based on these findings, others analyzed the impact of worker moods in struggling web search tasks [67]. Although the impact of worker moods on quality related task outcomes on traditional web interfaces is evident, how worker moods interact with conversational crowdsourcing to shape work quality in conversational crowdsourcing needs further investigation.

### Improving Worker Experience through Gamification

Gamification has been extensively used in the realm of crowdsourcing to make workers more motivated and engaged [154]. Following Flow theory [41], Eickhoff et al. designed a game to attract workers to execute Relevance Assessment tasks, resulting in lower cost and fewer malicious behaviors [55]. A prior work used competition-based design to improve worker performance in microtask crowdsourcing on the CrowdFlower (Appen) platform [183]. Furthermore, using gamification to exploit worker motivation and interest to enable volunteering crowdsourcing has shown to be feasible in prior studies [154]. A previous study developed an online collaborative game to effectively crowdsource protein structures [37]. Similar methods are also extensively used to inspire volunteers [18], increase enjoyment [130], or support activism [142]. Gamification has been shown to be effective in crowdsourcing. However, we learned that games from previous studies for crowdsourcing are all task-specific, meaning the game must be well designed to meet the requirements of simultaneously engaging workers and acquiring specific types of data. There are no common guidelines or tools for rapidly developing a game with little overhead.

***Identification with Avatars.*** Avatar customization is a simple interface manipulation and has shown to increase task engagement [15]. Avatars have been employed in many different areas, particularly in gaming systems. Prior work has showcased how and why players can be engaged in digital games [184], which is widely accepted by the researchers of relevant fields. The authors proposed self-determination theory (SDT) to explain the reason that games are usually engaging, and suggested that players would be intrinsically motivated if the game was designed to satisfy players' psychological needs of self-determination. Based on the model of enjoyment [225], Trepte et al. studied competitiveness, player life satisfaction, and avatar identification in video games. They found a strong relationship between avatar identification and game enjoyment [220]. Apart from the effect of game enjoyment, prior work found that avatar customization itself could be engaging and valuable to players, after the authors carried out an interview study about the game World of Warcraft [133]. Furthermore, giving personality traits or even names could be important to increase identification while creating an avatar [220, 146, 40]. Fictional characters or avatars sometimes present what users or players wish to be. Hoffner et al. interviewed children and young adults about their favorite characters. Results indicated that both similarity identification (gender) and wishful identification (characteristic) existed in their favorite characters [96, 97]. Furthermore, Neustaedter et al. presented a study showing players created and evolved the avatar in games to match a desired virtual identity [160]. The theories proposed and supported in previous works lead to a new research opportunity – to improve worker experience by enabling workers to customize their avatar appearance, and selection of their desired avatar characteristics.

## 1.2 Research Questions and Original Contributions

In the following, we present the research questions investigated in Chapters 2-5 and summarize the original contributions. As shown in Figure 1.2, we address research questions about designing conversational crowdsourcing (Chapter 2), using conversational crowdsourcing to improve worker satisfaction and engagement (Chapter 3), analyzing the roles of worker mood and self-identification in conversational crowdsourcing (Chapter 4), and applications of conversational crowdsourcing (Chapter 5). The links between the research questions and research background are shown in Figure 1.1. To answers these research questions, we carried out 7 online experiments, involving more than 2000 workers, on three crowdsourcing platforms (Amazon's Mechanical Turk, Figure Eight, and Prolific).

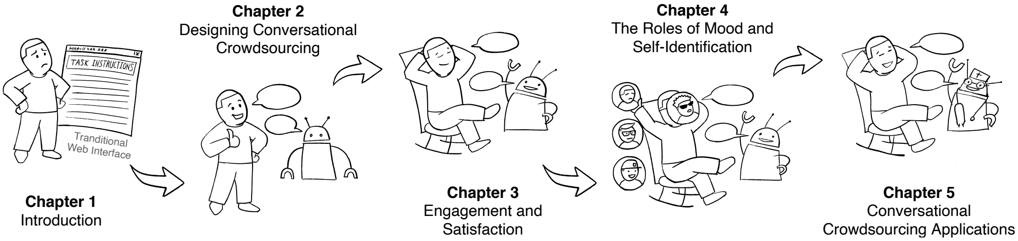

TASK INSTRUCTIONS
Tranditional Web Interface
Chapter 1
Introduction
Chapter 2
Designing Conversational Crowdsourcing
Chapter 3
Engagement and Satisfaction
Chapter 4
The Roles of Mood and Self-Identification
Chapter 5
Conversational Crowdsourcing Applications

Figure 1.2: The thesis outline.

### Designing Conversational Crowdsourcing

Traditional web-based user interfaces are widely used for the interaction between crowdsourcing platforms and workers in the majority of prior work, to communicate with workers, transmit instructions and gather responses thereafter. In Section 1.1, we have identified that conversational interfaces can improve user experiences. To demonstrate the feasibility of our novel conversational crowdsourcing paradigm, it is important to properly design the conversational interface in a way which contains all the essential task elements that traditional web interfaces have. Therefore, we aim to understand whether conversational crowdsourcing could result in reasonably high quality of output in comparison with traditional web crowdsourcing interfaces. To this end, in Chapter 2, we address the following research questions:

**RQ2.1:** To what extent can text-based conversational interfaces support the execution of different types of crowdsourced microtasks?

**RQ2.2:** How do different types of UI input elements in conversational interfaces affect quality-related outcomes in microtasks?

***Original contributions.*** This work takes an important first step of studying the utility of conversational interfaces in microtask crowdsourcing. The original contributions of Chapter 2 are threefold:

1. We designed the logic and workflow of the conversational agent for assisting crowd workers in task execution based on Telegram Bot (**RQ2.1**).

2. We show that tasks executed through conversational crowdsourcing interfaces can result in a similar output quality and execution time compared to traditional web interfaces (**RQ2.2**).

3. Considering the fact that most crowdsourcing platforms are web based, to further lower the barrier of participation, we developed a tool for quickly deploying crowdsourcing tasks in a customizable web-based conversational interface, named TickTalkTurk [173]. Our conversational crowdsourcing tasks and studies in the remaining chapters are primarily implemented based on TickTalkTurk.

### Improving Worker Engagement and Satisfaction

By addressing the above research question, we established conversational interfaces as a viable alternative to the existing standard web interfaces for microtask crowdsourcing. However, the impact of conversational crowdsourcing on the engagement of workers needs a thorough analysis. Tasks on crowdsourcing platforms are often deployed in large batches consisting of similar microtasks [3, 51]. The monotonous nature of crowdsourcing tasks can lead to sloppy work due to boredom and fatigue (Section 1.1), and then result in low worker satisfaction and engagement. Furthermore, previous work [156, 113] has shown that the design of the conversation can affect the worker experience, as previous psychological and linguistic studies have shown the important role that *conversational styles* can play in inter-human communication [125, 210, 211]. Therefore, it is important to investigate whether the insights and conclusions about conversational styles in human conversation are applicable to conversational crowdsourcing. In Chapter 3, we aim to address the following research question:

**RQ3.1:** To what extent can conversational crowdsourcing improve the worker engagement?

**RQ3.2:** How do different conversational styles affect the performance of workers and their cognitive load in conversational crowdsourcing?

Understanding the role of conversational styles of online workers while performing crowdsourcing tasks can help us better design strategies to improve output quality and worker satisfaction. While we can simply assign a conversational style to a conversational agent, for workers, however, there is a need for research focusing on the estimation of their conversational styles. Therefore, we will delve into the following research questions:

**RQ3.3:** How can the conversational style of a crowd worker be reliably estimated?

**RQ3.4:** To what extent does the conversational style of crowd workers relate to their work outcomes, perceived engagement, and cognitive task load in different types of tasks?

***Original contributions.*** Our findings have important implications on worker performance prediction, task scheduling and assignment in microtask crowdsourcing, and furthering the understanding of conversational crowdsourcing. Particularly, the original contributions of Chapter 3 are:

1. We show that conversational crowdsourcing have positive effects on worker engagement, as well as the perceived cognitive load, in comparison to traditional web-based crowdsourcing (**RQ3.1**).

2. We show that a High-Involvement conversational style can better engage workers for specific task types (**RQ3.2**).

3. Our results reveal that our novel coding scheme can estimate crowd workers' conversational styles with a high inter-rater reliability (**RQ3.3**).

4. We show that workers with an *Involvement* conversational style have significantly higher output quality, higher user engagement and less cognitive task load while they are completing a high-difficulty task, and have less task execution time in general (**RQ3.4**).

### The Roles of Worker Moods and Self-identification

Conversational crowdsourcing is associated with human emotions due to its human-like means of interaction [196]. Therefore, there is a research opportunity about the role of worker emotions in shaping work in conversational crowdsourcing. Recently, previous studies have shown evidence that worker moods can affect quality related outcomes and worker experience [244]. As we discussed before (Section 1.1), we lack an understanding of the effect of workers' subjective perceptions. Especially, the effect of moods in conversational crowdsourcing needs further investigation. Therefore, in Chapter 4, we will first try to address the following research questions:

**RQ4.1:** How do worker moods affect their performance, engagement and cognitive load in conversational crowdsourcing?

**RQ4.2:** How does the conversational style of a conversational agent affect the performance of workers in different moods?

In the previous section, we have explained that self-identification is found to be strongly associated with emotions in the realm of games [225, 220], while gaming research is becoming increasingly popular in crowdsourcing. Recent work has shown that self-identification with player avatars is effective in fostering interest, enjoyment, and other emotional aspects pertaining to intrinsic motivation [14]. Self-identification with avatars is already very common in video games but not essentially in the realm of crowdsourcing (and naturally conversational crowdsourcing). To fill the knowledge gap, we thereby delve into the following research questions:

**RQ4.3:** How do worker avatars affect worker experience and quality-related outcomes in traditional web and novel conversational interfaces?

**RQ4.4:** How can worker self-identification with their avatars be facilitated using avatar customization and worker characterization?

***Original contributions.*** The findings provide useful insights for future crowdsourcing task design, with the aim of improving worker satisfaction and alleviating cognitive task load. Specifically, the original contributions of Chapter 4 are:

1. We show that, in conversational crowdsourcing, workers in a pleasant mood generally exhibited a higher output quality, higher user engagement and less cognitive task load (**RQ4.1**).

2. We show that a suitable conversational style of the agent can have a significant impact on worker performance (High-Involvement style for pleasant workers and High-Considerate style for unpleasant workers) (**RQ4.2**).

3. Using avatar appearance customization in conversational crowdsourcing can effectively reduce workers' perceived workload and improve quality-related outcomes (**RQ4.3**).

4. The results reveal the occurrence of similarity (avatar appearance) and wishful (avatar characterization) identifications with worker avatars (**RQ4.4**).

### Applications of Conversational Crowdsourcing

In Chapter 5, we study the performance of the conversational crowdsourcing paradigm in the context of two application domains: information retrieval and health studies.

Crowdsourcing has become a crucial means in the realm of information retrieval for carrying out user studies to test the effectiveness of novel information retrieval systems. Considering the advantages of using conversational interfaces in crowdsourcing, it is important to study whether improved user engagement through conversational interfaces can lead to better memorability of information due to the fact that information overload is a problem many of us can relate to nowadays. Furthermore, we also consider note-taking feature in the user interface as previous studies have shown the effectiveness of using note-taking to improve memorability in the classroom. To this end, we aim to fill the knowledge gap by proposing novel conversational interfaces with note-taking features to improve human memorability during information search. The research questions of the first conversational crowdsourcing application are:

**RQ5.1:** How can human memorability of information consumed in informational web search sessions be improved?

**RQ5.2:** How does the use of text-based conversational interfaces and note-taking affect the search behavior of users?

To take crucial strides to the future of crowdsourcing, in the second application of conversational crowdsourcing, we explore the prevailing psychosocial and mental health of crowd workers, to have a better understanding of worker health and wellbeing across different crowdsourcing platforms. We design a 60-item survey including four main aspects, namely, background and working environment, ergonomics and physical health, psychosocial conditions and mental health, and workers' needs for improving their health. Since we show that conversational crowdsourcing is effective in improving worker engagement in long and monotonous tasks, we apply conversational crowdsourcing to assist workers in completing the survey. In the second application, we delve into the following research questions:

**RQ5.3:** What is the prevalent physical and mental health status of crowd workers in microtask crowdsourcing marketplaces?

**RQ5.4:** To what extent are healthcare interventions needed in crowdsourcing marketplaces? What are the preferred characteristics of such interventions from the perspective of workers?

***Original contributions.*** Our work has important implications in improving human memorability in information retrieval, and on task and workflow design that are centered around worker health on crowdsourcing platforms. Specifically, the main contributions of Chapter 5 are:

1. We show that conversational interfaces have the potential to augment long-term memorability (**RQ5.1**).
2. We show that users leveraging conversational interfaces exhibit a completely different behavior pattern compared to traditional web users. Such behaviors have been proved to be beneficial for human memorability by previous studies (**RQ5.2**).
3. We show that workers across different crowdsourcing platforms report similar health-related issues, but also exhibited certain differences. Mechanical Turk workers reported better physical health, while Prolific workers reported better mental health (**RQ5.3**).
4. We show that physical discomfort is related to the working ergonomics of crowd workers (**RQ5.3**).
5. We show that workers' energy levels could be affected by task content (the meaning of work and possibilities of learning); the mental wellbeing of crowd workers could be affected by their work pace and task demands (**RQ5.3**).
6. Our survey has shown that it would be appropriate to design and provide health interventions actively to workers (lasting no longer than 10 minutes), every 0.5-2 hours, between batches of tasks (**RQ5.4**).
7. According to workers' comments, the types of health interventions, their duration, content, and frequency should be customizable and personalized to worker preferences (**RQ5.4**).

## 1.3 Research Methodology

The work in this thesis can be categorized as *quantitative empirical user study.* We aim to provide analytical insights and suggest improvements by carrying out *controlled crowdsourcing experiments* on crowdsourcing platforms with the aim of studying worker performance and experience in crowd work. We collect and analyze data related to worker performance and worker behavior during task execution. The independent variables in our experiments include worker interfaces (addressing research questions in Chapter 2), conversational styles (addressing research questions in Chapter 3), and crowdsourcing tasks (all the research questions). The targeted participants are generic crowd workers. Since the goal is to study human factors in general crowd work, we do not set up any particular qualifications or filters to pre-screen crowd workers in our studies, except overall approval rate. Crowd workers in our experiments are mainly incentivized by monetary rewards as we aim to study worker performance and experience on paid crowdsourcing platforms. We apply quantitative data analysis after crowdsourcing experiments. To understand how worker performance and behavior relate to conversational crowdsourcing, we use statistical significance tests to verify our hypotheses and measure the reliability of our proposed methods. To answer the research questions, we carry out crowdsourcing experiments across all the chapters. We compare novel conversational crowdsourcing with traditional web-based crowdsourcing using different task types. We measure and analyze worker accuracy, task execution time, worker retention (research questions in Chapter 3 and Chapter 4), and worker memorability (research questions in Chapter 5).

Controlled crowdsourcing experiments provide us with objective data and properties (such as worker accuracy, worker retention, and task execution time) that we need for analyzing the performance and behavior of crowd workers. However, sensing crowd workers' subjective feelings and understanding why crowd workers exhibit certain behaviors are as equally important as measuring objective performance data in our studies. *Surveying crowd workers* allows us to acquire important data related to worker experience and satisfaction during task execution. Therefore, we survey crowd workers to acquire their perceived engagement, cognitive workload, and intrinsic motivation, after they complete all the crowdsourcing tasks. In this thesis, we survey crowd workers across all the chapters. In Chapter 2, we survey the workers who participate the crowdsourcing experiment, by asking them to optionally give overall satisfaction scores. In Chapter 3 and Chapter 4, the workers who participate the crowdsourcing experiments are asked to provide their ratings about perceived user engagement, cognitive task load, worker mood, and intrinsic motivation (only Section 4.2). In Chapter 5, we survey workers on Amazon Mechanical Turk and Prolific platforms. We design survey questionnaires to understand worker health in prevalent crowdsourcing marketplaces.

## 1.4 Origin of Chapters

This thesis consists of six chapters. The current chapter (**Chapter 1**) describes the research background, research questions, and original contributions. Chapters 2-5 are based on research papers published in conferences and journals:

**Chapter 2** is based on a full research paper published at the 2019 ACM Conference on User Modeling, Adaptation and Personalization [144] and a demonstration paper published at the 2020 ACM Conference on Computer-Supported Cooperative Work and Social Computing [173].

**Chapter 3** is based on two full research papers published at the 2020 ACM CHI Conference on Human Factors in Computing Systems [171] and the Proceedings of the ACM on Human-Computer Interaction (CSCW 2020) [170] respectively.

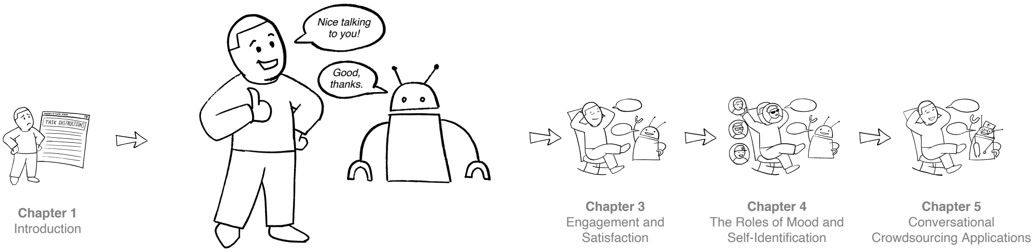
**Chapter 4** is based on two full research papers published at the International Conference on Web Engineering 2020 [172] and the Proceedings of the ACM on Human-Computer Interaction (CSCW 2021) [169] respectively.

**Chapter 5** is based on a full research paper published at the Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval [174] and a research paper currently in submission [175].

**Chapter 6** concludes the thesis by summarizing the main findings and contributions and providing an outlook to future research directions in related fields.

# Chapter 2

# Designing Conversational Crowdsourcing


TASK INSTRUCTIONS
Chapter 1
Introduction
Nice talking to you!
Good, thanks.
Chapter 3
Engagement and Satisfaction
Chapter 4
The Roles of Mood and Self-Identification
Chapter 5
Conversational Crowdsourcing Applications

Conversational interfaces can facilitate human-computer interactions. Whether or not conversational interfaces can improve worker experience and work quality in crowdsourcing marketplaces has remained hitherto unanswered. In this chapter, we investigate the suitability of text-based conversational interfaces for microtask crowdsourcing. We designed a rigorous experimental campaign aimed at gauging the interest and acceptance by crowd workers for this type of crowd work interface. We designed conversational interfaces for task execution based on the messaging application Telegram, compared web and conversational interfaces for five common microtask types, and measured the execution time, quality of work, and the perceived satisfaction of 316 workers recruited from the Figure Eight (Appen) platform. We show that conversational interfaces can be used effectively for crowdsourcing microtasks, resulting in a high satisfaction from workers, and without having a negative impact on task execution time or work quality.

The experimental campaign we carried out also exposed weaknesses of using a third-party messaging application independent of the crowdsourcing platform (e.g., Telegram), which required workers to put extra effort into installing the application and registering a new account in order to complete the task. To further lower the barrier for participation, we designed a web-based conversational crowdsourcing tool named TickTalkTurk, to assist task requesters in quickly deploying and publishing conversational microtask on popular crowdsourcing platforms.

The content of this chapter is based on the following papers:

Panagiotis Mavridis, Owen Huang, Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon. Chatterbox: Conversational Interfaces for Microtask Crowdsourcing. Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, pp. 243-251, 2019. (Section 2.1 is based on this paper. This paper is derived from a master thesis project supervised by Sihang Qiu. He was responsible for design and execution of the study, analysis of data, and paper-writing of the corresponding parts)

Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon. TickTalkTurk: Conversational Crowdsourcing Made Easy. Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing, pp.53-57, 2020. (Section 2.2 is based on this demonstration paper)

## 2.1 Conversational User Interface for Crowdsourcing

Messaging applications such as Telegram, Facebook Messenger, and Whatsapp, are regularly used by an increasing number of people, mainly for interpersonal communication and coordination purposes [132]. Users across cultures, demographics, and technological platforms are now familiar with their minimalist interfaces and functionality. Such popularity, combined with recent advances in machine learning capabilities, has spurred a renewed interest in conversational interfaces [245], and *chatbots*, i.e. text-based conversational agents that mimic a conversation with a real human to enable conversational, information seeking [5, 6, 226], and transactional tasks [53, 65, 246].

The growing popularity of conversational interfaces has coincided with flourishing crowdsourcing marketplaces. Microtask crowdsourcing allows the interaction with a large crowd of diverse people for data processing or analysis purposes. Examples of such microtasks include audio/text transcription, image/text classification, and information finding. Microtask crowdsourcing is commonly executed by means of dedicated web platforms (e.g. Amazon Mechanical Turk, Appen), where all the published microtasks are publicly presented to workers. Upon the selection of their preferred microtasks, workers are typically directed to a webpage served by the platform or hosted on an external server by the task requesters. Based on the task design, workers can provide their input by means of standard (e.g. text, dropdown, and multiple choice fields) or custom (e.g. drawing tools) web UI elements. Recent work has shed light on the importance of task design choices made with respect to user interface elements; and on how such choices can influence the quality of work produced and satisfaction among workers [66].

Lowering the entry barrier for workers to participate effectively in crowdsourcing tasks is an important step towards securing the future of crowd work [117]. Messaging applications using conversational interfaces are reported to be more popular than social networks [201], and we argue that such familiarity with conversational interfaces can potentially lower the barrier for participation. Although conversational interfaces have been effectively used in numerous applications, the suitability and effectiveness of conversational interfaces in microtask crowdsourcing marketplaces has remained unexplored. We aim to address this knowledge gap in this section. We investigate the suitability of conversational interfaces for microtask crowdsourcing by juxtaposing them with standard web interfaces in a variety of popularly crowdsourced tasks.

Our goal is to further the understanding of how text-based conversational interfaces could serve as an alternative to the standard web interfaces typically used for microtask crowdsourcing. We seek answer to the following questions:

**RQ2.1:** To what extent can text-based conversational interfaces support the execution of different types of crowdsourced microtasks?

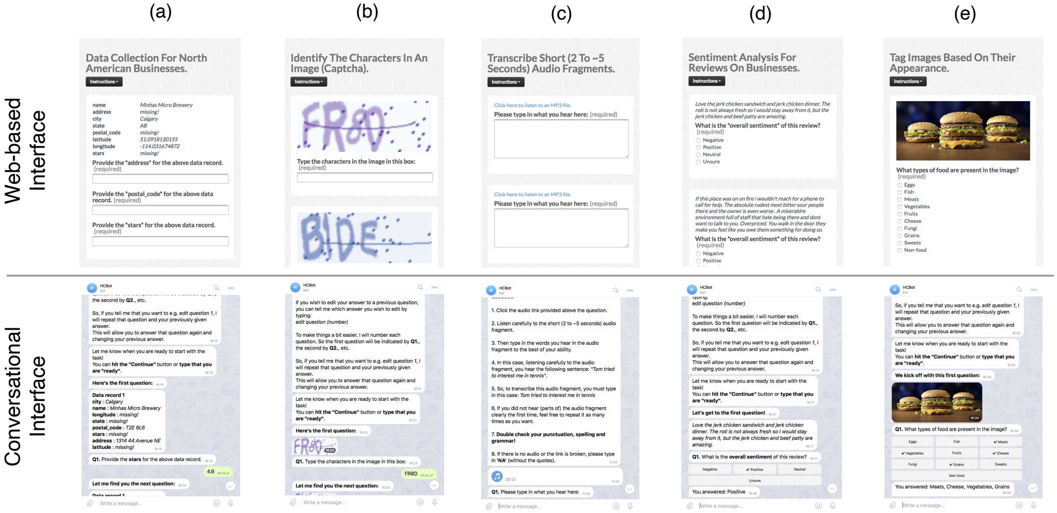
**RQ2.2:** How do different types of UI input elements in conversational interfaces affect quality-related outcomes in microtasks?

We carried out experiments to gauge the interest and acceptance of automated, text-based conversational work interfaces by crowd workers, while assessing their performance within different task types. We recruited workers from the Figure Eight (Appen) microwork platform, and implemented a conversational interface based on the popular *Telegram* messaging platform. We addressed five typical microtask types (information finding, human OCR (captcha), speech transcription, sentiment analysis, image annotation) spanning content types (text, image, audio) and UI elements (free text, single and multiple selections). For each task type, we implemented both web and conversational interfaces.

We addressed **RQ2.1** by comparing the execution time, quality of results, and satisfaction of workers who used the standard web interface with those who used the conversational interface. To answer **RQ2.2**, we compared different implementations of conversational UI elements for single and multiple input selections in microtasks. Results showed that the conversational interfaces were positively received by crowd workers, who indicated an overall satisfaction and an intention for future use of similar interfaces. In terms of performance, tasks executed using the conversational interfaces took similar execution times, and yielded comparable output quality.

## Study Design

We considered five types of microtasks that are typically completed by crowd workers in microwork crowdsourcing marketplaces. We selected these tasks both to stress the diversity of evaluated content types (text, images, audio), and the diversity of UI elements used to perform the tasks. For the sake of reproducibility, the complete list of tasks (and related data) is available for download on the companion webpage.[5]



Data Collection For North American Businesses.

name Minhas Micro Brewery
address missing!
city Calgary
state AB
postal_code missing!
latitude 51.0918130155
longitude -114.031674872
stars missing!

Provide the *address* for the above data record. (required)

Provide the *postal_code* for the above data record. (required)

Provide the *stars* for the above data record. (required)

Identify The Characters In An Image (Captcha).

Type the characters in the image in this box: (required)

Transcribe Short (2 To ~5 Seconds) Audio Fragments.

Click here to listen to an MP3 file.
Please type in what you hear here: (required)

Click here to listen to an MP3 file.
Please type in what you hear here: (required)

Sentiment Analysis For Reviews On Businesses.

*Love the jerk chicken sandwich and jerk chicken dinner. The roti is not always fresh so I would stay away from it, but the jerk chicken and beef patty are amazing.*
What is the *overall sentiment* of this review? (required)
○ Negative
○ Positive
○ Neutral
○ Unsure

*If this place was on on fire i wouldn't reach for a phone to call for help. The absolute rudest most bitter sour people there and the owner is even worse . A miserablre environment full of staff that hate being there and dont want to talk to you. Overpriced. You walk in the door they make you feel like you owe them something for doing so.*
What is the *overall sentiment* of this review? (required)
○ Negative
○ Positive

Tag Images Based On Their Appearance.

What types of food are present in the image? (required)
□ Eggs
□ Fish
□ Meats
□ Vegetables
□ Fruits
□ Cheese
□ Fungi
□ Grains
□ Sweets
□ Non-food

the second by **Q2.**, etc.

So, if you tell me that you want to e.g. *edit question 1*, I will repeat that question and your previously given answer.
This will allow you to answer that question again and changing your previous answer.

Let me know when you are ready to start with the task!
You can **hit the "Continue"** button or **type that you are "ready".**

**Here's the first question:**

**Data record 1**
**city** : *Calgary*
**name** : *Minhas Micro Brewery*
**longitude** : *missing!*
**state** : *missing!*
**postal_code** : *T2E 6L6*
**stars** : *missing!*
**address** : *1314 44 Avenue NE*
**latitude** : *missing!*

**Q1.** Provide the **stars** for the above data record.

4.6

**Let me find you the next question:**

If you wish to edit your answer to a previous question, you can tell me which answer you wish to edit by typing:
*edit question (number)*

To make things a bit easier, I will number each question. So the first question will be indicated by **Q1.**, the second by **Q2.**, etc.

So, if you tell me that you want to e.g. *edit question 1*, I will repeat that question and your previously given answer.
This will allow you to answer that question again and changing your previous answer.

Let me know when you are ready to start with the task!
You can **hit the "Continue"** button or **type that you are "ready".**

**Here's the first question:**

**Q1.** Type the characters in the image in this box:

FR8D

**Let me find you the next question:**

1. Click the audio link provided above the question.

2. Listen carefully to the short (2 to ~5 seconds) audio fragment.

3. Then type in the words you hear in the audio fragment to the best of your ability.

4. In this case, listening carefully to the audio fragment, you hear the following sentence: "*Tom tried to interest me in tennis*".

5. So, to transcribe this audio fragment, you must type in this case: *Tom tried to interest me in tennis*

6. If you did not hear (parts of) the audio fragment clearly the first time, feel free to repeat it as many times as you want.

7. **Double check your punctuation, spelling and grammar!**

8. If there is no audio or the link is broken, please type in **'NA'** (without the quotes).

00:01

**Q1.** Please type in what you hear here:

*edit question (number)*

To make things a bit easier, I will number each question. So the first question will be indicated by **Q1.**, the second by **Q2.**, etc.

So, if you tell me that you want to e.g. *edit question 1*, I will repeat that question and your previously given answer.
This will allow you to answer that question again and changing your previous answer.

Let me know when you are ready to start with the task!
You can **hit the "Continue"** button or **type that you are "ready".**

**Let's get to the first question!**

*Love the jerk chicken sandwich and jerk chicken dinner. The roti is not always fresh so I would stay away from it, but the jerk chicken and beef patty are amazing.*

**Q1.** What is the **overall sentiment** of this review?

Negative | ✓ Positive | Neutral
Unsure

You answered: Positive

So, if you tell me that you want to e.g. *edit question 1*, I will repeat that question and your previously given answer.
This will allow you to answer that question again and changing your previous answer.

Let me know when you are ready to start with the task!
You can **hit the "Continue"** button or **type that you are "ready".**

**We kick off with this first question:**

**Q1.** What types of food are present in the image?

Eggs | Fish | ✓ Meats
✓ Vegetables | Fruits | ✓ Cheese
Fungi | ✓ Grains | Sweets
Non-food

You answered: Meats, Cheese, Vegetables, Grains

Figure 2.1: In this figure we depict different tasks (a, b, c, d, e) and how they look from a Standard web (top) versus a conversational (bottom) interface perspective. The different types of tasks depicted: a) Information Finding, b) Human OCR, c) Speech Transcription, d) Sentiment Analysis, e) Image Annotation.

*Information Finding.* Workers are tasked to find specific relevant information from a given data source [72]. We opted for business-related information available on the Web, to facilitate

[5] https://sites.google.com/view/umap2019chatbotmicrowork

retrieval and minimize task execution delays due to hard-to-find information. We used the first 17 business records listed in the `Yelp dataset`[6]. From these 17 records, we created 50 task objects by randomly removing three of the following fields: *name*, *address*, *city*, *state*, *postal code* and *stars* (i.e. the business rating). To prevent ambiguity, the *name* and *postal code* were never jointly removed from the same business record. The workers' task was to use commercial search engines to retrieve the missing information from the business record, and to provide it as free text in three separate fields.

*Human OCR (CAPTCHA).* This is a media transcription task [72], where workers were required to transcribe the text contained in a CAPTCHA image. We generated[7] 50 distinct CAPTCHAs of four characters, containing only digits and letters (i.e. excluding special characters and symbols such as punctuation marks, currency symbols, etc.).

*Speech Transcription.* In this audio transcription task, workers were asked to transcribe recordings of English speech retrieved from Tatoeba[8]. We selected 50 distinct recordings, with length ranging from 2 to 8 seconds, and asked workers to type the content of the short speech.

*Sentiment Analysis.* In this task, workers were asked to assess the sentiment of user reviews. We relied again on the `Yelp` dataset, and selected 50 reviews. To maintain sufficient diversity on selected businesses, we selected a maximum of three reviews per business. The length of the selected reviews varied, ranging from several sentences to whole paragraphs. Workers were asked to judge the *overall sentiment* of a review as *Positive*, *Negative*, or *Neutral*. An additional *Unsure* option was provided, to address annotation uncertainty and prevent forced choices.

*Image Annotation.* This is another data enhancement task where the goal is to determine the categories of the food items contained in an image. The options included: *Eggs*, *Fish*, *Meat*, *Vegetables*, *Fruits*, *Cheese*, *Mushroom*, *Grain*, and *Sweets*. In case the image did not contain any food category that was applicable, workers were requested to only select a *Non-food* option. We used 50 distinct images from the `Yelp` dataset.

### Work Interfaces

We focused on three types of UI elements that are required to perform the task types investigated in our experiments as shown in Table 2.1; (1) *Free Text*, to input text data retrieved from the Web, annotations about a data object, or transcriptions from images and sound; (2) *Single Selection from List*, for single-class classification (*Sentiment Analysis*); and (3) *Multiple Selection from List*, for multi-class classification (*Image Annotation*).

The following sections describe and justify the interface designs adopted in our work. All the implemented interfaces are available on the companion webpage for reference.

***Standard Web Interface.*** The web interface was developed on the Figure Eight platform, which provides a standardized way to specify work interfaces in an HTML-like format. We decided to use only standard interface elements, that are typical of crowdsourcing tasks on Figure Eight, to elicit normal interactions of workers with the web interface.

Figure 2.1 depicts a one-to-one comparison of the Standard Web Interface tasks versus the Conversational Interface tasks.

[6] Yelp dataset: `https://www.yelp.com/dataset`

[7] CAPTCHA generator: `https://pypi.org/project/captcha/`

[8] `https://tatoeba.org/eng/audio/index`

Table 2.1: Summary of considered UI elements, and their implementation in web and conversational interfaces.

| ***UI Element*** | **Web** | **Conversational** |
|---|---|---|
| *Free Text* | Single/Multi line text | Message |
| *Single Selection* | Radio buttons | Single Button |
| *Multiple Selection* | Checkbox(es) | Multiple Buttons |

We can see the screenshots of the developed Web UIs corresponding to each of the 5 task types. Figure Eight provides two types of *Free Text* UI elements: `single line` text input and `multi-line` text input. The former type is used in the *Information Finding* and *Human OCR* tasks, as worker were asked to provide short input text (e.g. business name, city, address). The latter type is used in the *Speech Transcription* task, workers had to input short sentences from the processed audio. The *Single Selection* element needed for the *Sentiment Analysis* task has been implemented using `Radio Buttons`, as customary for this type of tasks; while the *Image Annotation* tasks used the `Checkboxes` UI element for *Multiple Selection.* When the task entailed multiple annotations (e.g. sentiment analysis, image labeling), content items and their respective input elements were presented in a sequence, to be navigated top-to-bottom within the same page.

***Conversational Interface.*** To resonate with popular conversational interfaces, we designed and implemented our conversational interface in the `Telegram`[9] messaging platform.

The interface comprises two main modules: 1) a *conversation management* module, responsible for aligning the status of the task execution with the status of a conversation, and for supporting navigation within the conversation ; and 2) an *input management* module, responsible for rendering the content associated to a task, and the UI elements required to allow and control user input.

Microtask crowdsourcing user interfaces are typically designed to be minimalistic and easy to use, to enable fast and effective work execution [117]. We shared the same design principle in the creation of the *conversation management* module, which consists of five simple states as illustrated in Figure 2.2. Figure 2.3 shows a brief example of the conversational flow in the chat interface.

***1)*** At the beginning of the task execution stage, a *chatbot* that drives the conversation, prompts the worker with messages containing task instructions, including an explanation of the task at hand, and examples of how input could be provided. ***2)*** Once no more annotations are pending in the task, the chatbot prompts the next question to the worker (content plus UI elements), and waits for the worker's response. ***3)*** Next, the answer provided by the worker is validated, with positive feedback if the answer is acceptable, or a re-submission sequence if the answer not valid. ***4)*** When no more annotations are pending, workers are shown their answers for review; and can ***5)*** re-process a previously submitted answer.

The *input management* component is built upon the standard *message* UI element, used by the workers and the chatbot to exchange information. Traditional text messaging systems only allow for alphanumeric content to be exchanged and rendered.

[9] `https://core.telegram.org/bots`

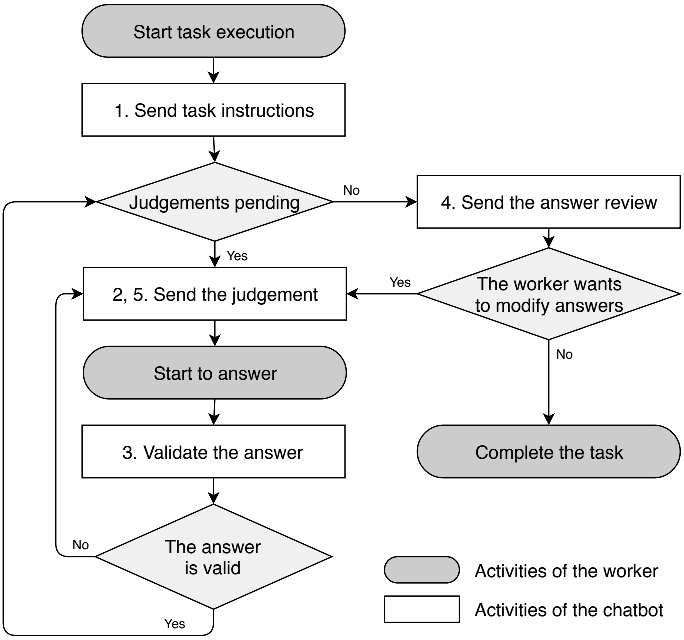
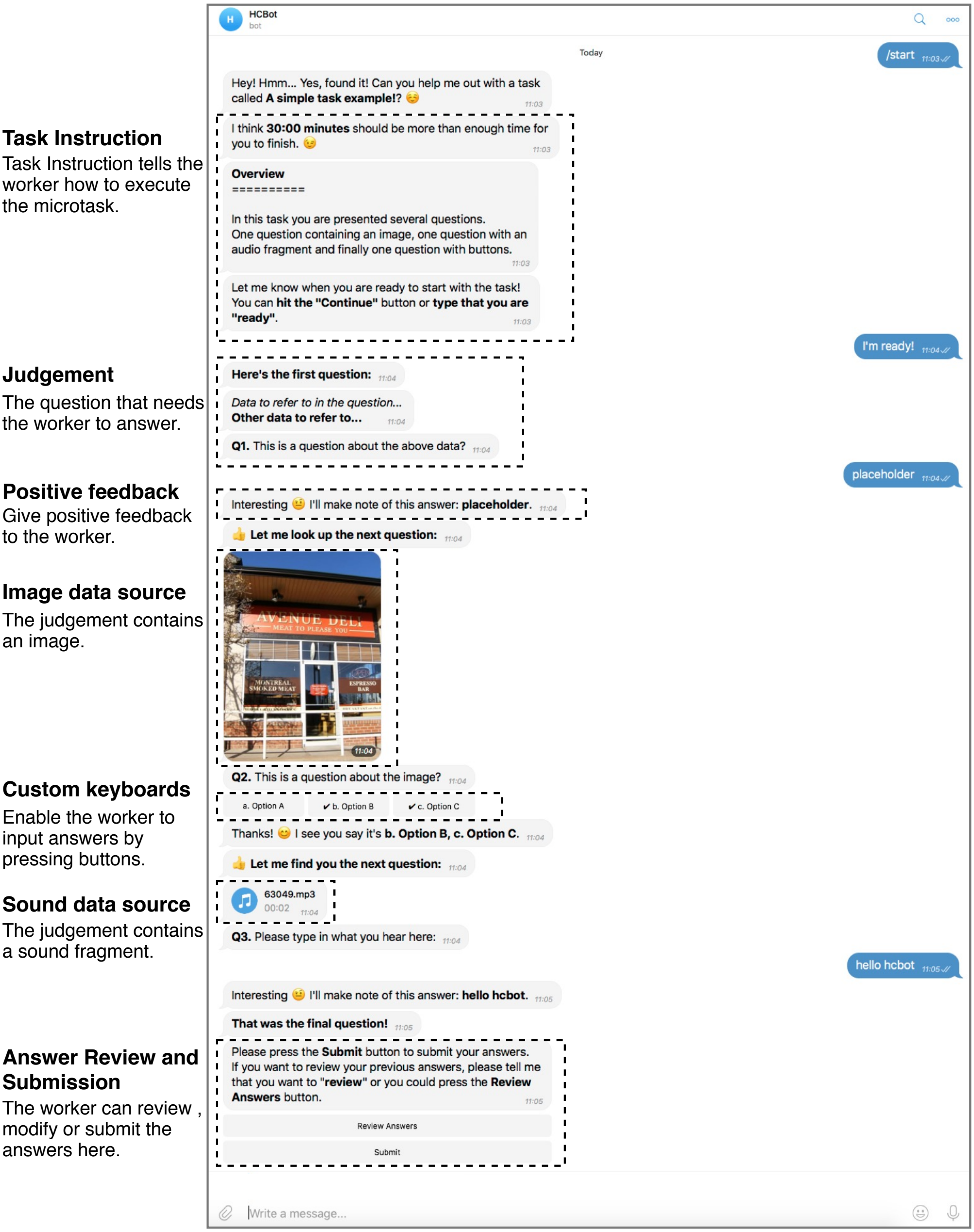


Figure 2.2: Conversation management logic.

Systems like *Telegram* allow for richer content, which include: *1) multimedia* content (images, videos, sound). *2) Interactive applications* (e.g. games), hosted on third party servers but rendered and accessible within the messaging application. *3) Custom keyboards*, which show predefined inputs, rendered textually or visually; notice that keyboards are complementary to the standard *message* element: the user can also simply type an abbreviated input (a single alphabet letter) used as a code associated with a pre-defined key option. And *4) commands*, i.e. instructions sent by the user to change the state of the chatbot (e.g. to start a new working session, or end an existing one).

Figure 2.1 depicts screenshots of the developed conversational interfaces. The design of both the interfaces and the interaction flows for each task type has been iterated and validated several times by the authors, through experiments with researchers and students from the research group. The *information finding* (a), *human OCR* (b), and *speech transcription* tasks (c) use a simple *message* element, where validation is performed by simply rejecting empty answers. The *sentiment analysis* (d) and *image annotation* (e) tasks were implemented with custom keyboards, allowing for (respectively) the single or multiple selection of predefined answers rendered as buttons associated with some option codes. Here, validation is performed by ensuring that only one button, option code, or content corresponding to an option is given. With custom keywords, workers could express their preference textually (with answers separated by whitespace or commas), using the option codes associated with the button, or by pressing the buttons. We use 4 custom keyboards configurations: 1) *Button-only Custom Keyboard*: Worker can select any button provided; 2) *Text-Only Custom Keyboard*: Worker can only type to provide its answer; 3) *Code-Only Custom Keyboard*: Worker can only type a letter to provide the answer from a predefined list; and 4)

*Mixed Custom Keyboard*: a worker can either select a button, type the full answer or the abbreviated code that corresponds to the answer (a single letter).



HCBot
bot

Today

/start 11:03

Hey! Hmm... Yes, found it! Can you help me out with a task called **A simple task example!?** ☺ 11:03

**Task Instruction**
Task Instruction tells the worker how to execute the microtask.

I think **30:00 minutes** should be more than enough time for you to finish. 😉 11:03

**Overview**
==========

In this task you are presented several questions.
One question containing an image, one question with an audio fragment and finally one question with buttons. 11:03

Let me know when you are ready to start with the task! You can **hit the "Continue"** button or **type that you are "ready"**. 11:03

I'm ready! 11:04

**Judgement**
The question that needs the worker to answer.

**Here's the first question:** 11:04

*Data to refer to in the question...*
**Other data to refer to...** 11:04

**Q1.** This is a question about the above data? 11:04

placeholder 11:04

**Positive feedback**
Give positive feedback to the worker.

Interesting 😀 I'll make note of this answer: **placeholder.** 11:04

👍 **Let me look up the next question:** 11:04

**Image data source**
The judgement contains an image.

11:04

**Q2.** This is a question about the image? 11:04

**Custom keyboards**
Enable the worker to input answers by pressing buttons.

a. Option A | ✔ b. Option B | ✔ c. Option C

Thanks! 😊 I see you say it's **b. Option B, c. Option C.** 11:04

👍 **Let me find you the next question:** 11:04

**Sound data source**
The judgement contains a sound fragment.

63049.mp3
00:02 11:04

**Q3.** Please type in what you hear here: 11:04

hello hcbot 11:05

Interesting 😀 I'll make note of this answer: **hello hcbot.** 11:05

**That was the final question!** 11:05

**Answer Review and Submission**
The worker can review , modify or submit the answers here.

Please press the **Submit** button to submit your answers. If you want to review your previous answers, please tell me that you want to **"review"** or you could press the **Review Answers** button. 11:05

Figure 2.3: An example showing the conversational interface developed for our experimental study.

In all the tasks types that we considered, the chatbot prompts the worker with the item to evaluate by rendering text (the business record to complete), images (the CAPTCHA and the food image), or speech (the audio to transcribe).

### Experimental Conditions

To answer **RQ2.1**, we designed 12 experimental conditions, with working interface type (Web, Conversational) and task type as independent variables, and the *Mixed Custom Keyboard* configuration for the *Sentiment Analysis* and *Image Labeling* conversational interfaces. As observable from Figure 2.1, the instructions at the beginning of the conversational task are relatively long, thus possibly affecting the task execution time.

To account for this, we include 6 additional experimental conditions where the conversational interface has task instructions partially hidden (workers are only presented with a brief overview of the task), and workers could instruct the chatbot through specific commands to display more detailed instructions (i.e. an example and its steps, and also inquire about how to edit a previously given answer). With **RQ2.2**, we tested the 3 *Custom Keyboard* configurations with the *Sentiment Analysis* and *Image Annotation* tasks, thus adding 6 additional experimental conditions.

### Task Assignment and Execution

On Figure Eight (Appen), we set up two types of jobs: *Web* jobs and *Conversational* jobs, where the latter included the string `*|*Requires Telegram*|*` in their title, to suggest the presence of a technical requirement for their execution.

Web jobs were completely performed within the Figure Eight platform, with the standard Figure Eight workflow and task assignment strategy.

Conversational jobs had a different flow: upon job selection, workers were informed that logging into Telegram was a requirement for participation. Additional instructions on how to register a Telegram account (if necessary) were also provided on an external web-page through a link. Several preview images were provided to inform workers about the nature of the task, and a short survey inquired about their working platform. We did not employ fingerprinting techniques to detect the digital work environment of workers to preserve worker privacy. Workers were informed that no personal information (e.g. names or phone numbers) would be stored, and that they would be allowed to withdraw from the experiment at any point in time.

To facilitate the assignment of tasks in Telegram, we redirect users via a URL to Telegram. According to their working environment, the worker could 1) have been redirect to a Web client version of Telegram; or, if the worker had a native Telegram client installed, 2) to the native Telegram application. Task assignment was performed dynamically, with a round robin policy on the content to be processed. A click of the *Submit* button commanded the finalization of the task, which resulted in a randomly generated validation token to be used in Figure Eight to fully complete the task and receive payment. Workers were also asked to indicate their intention to perform a similar task again in Telegram (yes/no)[10], and to optionally provide a comment about their working experience.

[10] *Would you be interested in doing a similar task again in Telegram?*

### Evaluation Metrics

The dependent variables in our experiment are *Execution Time*, *Answer Quality*, and *Workers Satisfaction.* Ground truth and evaluation data is available on the companion Web page.

*Execution Time* is measured as the time (in seconds) between the start and the submission of a task. In the web interface, this is calculated as the time from when the Figure Eight task is initiated, up to the moment the *Submit* button is clicked. In the conversational interface, this is calculated as the time difference between a click event on the *Start* button, and a click event on the *Submit* button.

*Answer quality* is measured by comparing the worker answers with ground truth *Sentiment Analysis* and *Image Annotation.* For the *Information Finding* and *Speech Transcription* task, workers results were manually inspected by the authors; simple syntactical and grammatical errors were tolerated. For the *human OCR* task, we compared the entire answer to the label of the CAPTCHA, disregarding errors with capitalization. To judge whether a worker had answered correctly for the *Image Annotation* task, we marked an answer as correct, as long as it contained at least one correct annotation, and no more than two wrong annotations.

*Workers Satisfaction* of both web and chatbot tasks is measured by default task ratings on Figure Eight (workers will be re-directed back to Figure Eight when they submit the answer on Telegram) after workers finish the task. Furthermore, for the chatbot tasks, the optional comments are left at the end of the chatbot task to let workers give their personal opinions.

## Evaluation

The experiments were performed recruiting workers from the Figure Eight microtask crowdsourcing platform. As the main objective of this work is to understand if text-based conversational agents can enable microtask crowdsourcing, we did not condition the participation of workers to pre-existing quality levels, nor did we run qualification tests. Each experimental condition has been deployed as a separate job in Figure Eight (Figure Eight). Each job contained 50 task instances, totaling 1200 executions for the whole experiments. Each instance has been compensated $0.15USD$. *Information Findings* tasks contained 1 business record; *Human OCR* tasks contained 5 distinct CAPTCHAs, *Speech Transcription* tasks contained 3 audio samples; *Sentiment Analysis* tasks contained 3 reviews; *Image Annotation* tasks contained 3 images each. The distribution and frequency of objects in Web and Chatbot tasks were identical. Workers could only execute one task instance per available job. Web and Chatbot jobs were deployed on different dates, to maximize the chance of obtaining disjoint worker populations. The statistical tests that we performed to test the significance are always Mann-Whitney-Wilcoxon pair-wise significance test.

316 distinct workers executed at least one task ($\mu = 3.886$, $\sigma = 2.4941$, $median = 2$). 31 workers executed both web and chatbot jobs. 12.2% of the workers self-reported that they performed chatbot jobs with a mobile device. To eliminate the influence of malicious behavior, a manual inspection of workers' submissions was conducted. Consequently, 19 workers are excluded in web tasks, and 33 workers are excluded in chatbot tasks.

### Standard Web versus Conversational Interfaces

**Execution Time**. Table 2.2 and Figure 2.4 depict basic statistics and the distribution of execution times for the considered experimental conditions. With the exception of the *Human OCR* task and the *Sentiment Analysis* task, the execution time distributions for the specific task types have no statistically significant difference (Mann-Whitney-Wilcoxon pair-wise significance test, region of rejection $p > 0.05$). *Speech Transcription* tasks show a slightly longer execution time, a result that we account to the UI design of the Web task, which, by forcing workers to open another browser tab to play the audio sample, might have caused delays.

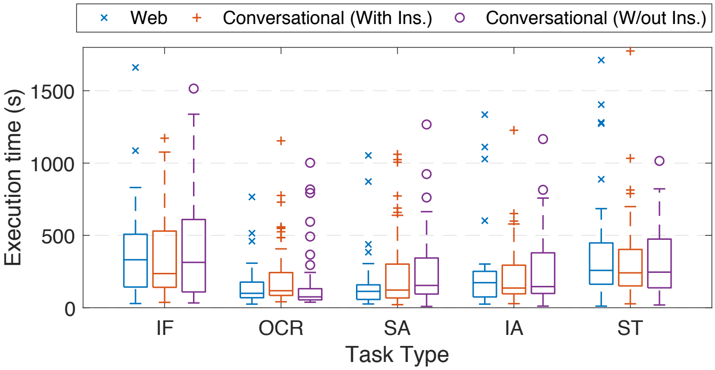


Figure 2.4: Tasks execution time (sec): Web vs. Conversational with instructions vs. Conversational without instructions.

Table 2.2: Execution time ($\mu \pm \sigma$: average and standard deviation, unit: seconds) in each work interface. With Ins.: with instructions; W/out Ins.: without instructions.

| ***Task type*** | **Web** | **Conversational** | |
|---|---|---|---|
| | | With Ins. | W/out Ins. |
| *Information Finding* | $364 \pm 301$ | $362 \pm 295$ | $393 \pm 328$ |
| *Human OCR* | $150 \pm 135$ | $219 \pm 227$ | $160 \pm 209$ |
| *Speech Transcription* | $384 \pm 381$ | $333 \pm 306$ | $311 \pm 223$ |
| *Sentiment Analysis* | $158 \pm 187$ | $243 \pm 276$ | $244 \pm 247$ |
| *Image Annotation* | $223 \pm 264$ | $222 \pm 212$ | $261 \pm 249$ |

The statistically significant difference between the *Sentiment Analysis* tasks (web vs. chatbot without instructions, $p = 0.03$) and the *Human OCR* tasks (chatbot with instruction vs. chatbot without instructions, $p = 0.01$) could be explained by the presence of long textual instructions at the beginning of the conversational interface which, differently from the Web interface, could not be hidden. This hypothesis is supported by the results obtained with the chatbot configurations where instructions were not initially visible: for all task types, execution time are lower, and with no statistical difference from their Web counterpart. Interestingly, only within very few tasks (10) workers executed the chatbot command to fully display task instructions, but in 150 occasions they asked to instructions steps or instructions examples at the beginning of the task. Finally, it is worth mentioning that in 84 occasions workers used the task reviewing and editing functionality, to correct their answers before submitting the results.

**Work Quality**. Table 2.3 summarizes the work performance evaluation for the considered task types. We observe comparable performance across tasks, with precision that is slightly lower (on average) with chatbot tasks. A manual analysis of the results highlights and interesting difference with the *Human OCR* tasks, where errors were mostly due to ambiguous characters in the CAPTCHA (e.g. "D" looking like either a capital "O" or a "0" (zero), rotated "L" looking like a "V"), but less present with chatbot workers. An analysis of the reasons beyond this result is left to future work.

Table 2.3: Quality of crowdwork produced across different task and interface types.

| ***Task type*** | **Web** | **Conversational** |
|---|---|---|
| *Information Finding* | 0.95 | 0.92 |
| *Human OCR* | 0.75 | 0.82 |
| *Speech transcription* | 0.85 | 0.75 |
| *Sentiment analysis* | 0.93 | 0.88 |
| *Image annotation* | 0.90 | 0.81 |

**Workers Satisfaction**. Workers participating in Chatbot tasks were also asked to provide feedback on their experience with microwork executed through conversational interfaces. 349 out of 600 executions received comment. Workers reported a positive opinion in 81.9% of comments. 44 workers gave a neutral comment. 19 workers indicated the issue about the slow response of the conversational interface.

The majority of the comments highlighted the intuitive user experience (e.g. *"Very easy to understand , and easy and fastest now we have buttons"*, *"very pleasant experience, i like the replays from the BOT, very interactive! Thx!"*, *"i loved this task, is so much different to the others, and i think is a excellent work it with telegram. nice"*, *"It was different, but i like it.."*, *"Yeah, i like this type of Task, is cool, a new feature is coming to us"*). Others remarked the enjoyable experience (*"This is fun and easy task I may try another task like this! Great!"*, *"Its fun!! best experience for first time using telegram haha"*). Some workers reported issues with the *"complicated"* set up, or with instructions that could be improved (*"MEJORAR LAS INSTRUCCIONES"* – *"Improve the instructions"*).

Table 2.4 reports the average Overall (**OV**), Instruction (**IN**), Ease of Job (**EA**), and Pay (**PA**) ratings given by workers after finishing the tasks. These ratings, expressed in a range between 1 and 5, are requested by the Figure Eight platform, and are optionally provided by workers. Ratings for Standard Web interfaces are to be considered as references for the deployed task types and object instances. Conversational interfaces received on average high, although slightly lower ratings than the ones received by Web interfaces. The difference is evident especially with the *Information Finding* task, where workers reported significantly lower ratings for all considered dimensions. With *Sentiment Analysis* tasks, ratings highlight differences in instructions and ease of use. With *Human OCR*, *Image Annotation*, and *Speech Transcription* ratings are comparable.

Table 2.4: Ratings of workers satisfaction. OV: Overall; IN: Instruction; EA: Ease of Job; PA: Pay

| ***Task type*** | **Platform** | **OV** | **IN** | **EA** | **PA** |
|---|---|---|---|---|---|
| *Information Finding* | Web | 4.5 | 4.3 | 4.5 | 4.5 |
| | Conversational | 3.0 | 3.4 | 2.9 | 3.3 |
| *Human OCR* | Web | 4.3 | 4.2 | 4.0 | 4.5 |
| | Conversational | 3.4 | 4.0 | 3.8 | 4.3 |
| *Speech Transcription* | Web | 4.7 | 4.7 | 3.9 | 4.1 |
| | Conversational | 4.1 | 4.5 | 4.0 | 4.3 |
| *Sentiment Analysis* | Web | 4.3 | 4.3 | 4.1 | 4.1 |
| | Conversational | 3.7 | 3.4 | 3.2 | 3.8 |
| *Image Annotation* | Web | 3.8 | 3.8 | 3.3 | 3.8 |
| | Conversational | 3.7 | 3.9 | 3.1 | 3.9 |

### Conversational Interfaces — UI Elements

Figure 2.5 and Table 2.5 depict basic statistics and the distribution of execution times for the considered experimental conditions. The use of different custom keyboards have an impact on the task execution times, both for single- and multiple-selection tasks, with statistically significant differences (Mann-Whitney-Wilcoxon pair-wise significance test, $p < 0.05$) with the text configuration ($p = 0.0011$ for *Sentiment Analysis* and $p = 0.0036$ for *Image Annotation*) and the code configuration ($p = 0.0003$ for *Sentiment Analysis*).

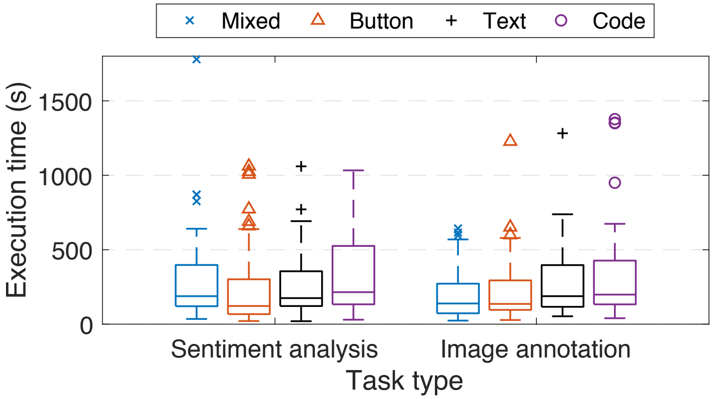


Figure 2.5: Task execution time (in seconds) with different *custom keyboard* configurations.

For the multiple-selection tasks, the availability of multiple input alternatives (*Mixed Custom Keyboard*) yields faster execution times; however, no clear total order of performance emerge across the two tasks. The removal of button shortcuts has a detrimental effect on workers execution time, while output quality is not affected. This is due to the input validation mechanism implemented in the conversational interface, that prevents wrong results from being submitted.

Table 2.5: Execution time ($\mu \pm \sigma$: average and standard deviation, unit: seconds) in each conversational interface. The *Mixed* configuration is the one adopted in **RQ2.1** experiments.

| ***Task type*** | **Mixed** | **Button** | **Text** | **Code** |
|---|---|---|---|---|
| *Sentiment Analysis* | $301 \pm 306$ | $243 \pm 276$ | $325 \pm 257$ | $267 \pm 219$ |
| *Image Annotation* | $211 \pm 178$ | $222 \pm 212$ | $339 \pm 342$ | $284 \pm 233$ |

### Design Implications

Results showed that the conversational interface could be a suitable alternative to Web-based microwork platforms, at least for the considered task types, both in terms of execution time and quality. As highlighted by previous work in mobile crowdsourcing [124, 45, 121], task and interaction design matter. Results suggest that for common tasks like *Sentiment Analysis* and *Image Labeling*, custom keyboard can enable execution times comparable to Web interfaces. Instructions and chatbot commands also have an impact, especially for domain specific tasks (e.g. food labeling).

Workers expressed positive opinions about this work interface modality. The analysis of workers' satisfaction highlight some differences across task types. While execution time and quality of output are comparable, workers were less satisfied with the quality of the instructions and ease of job (*Information Finding*, *Sentiment Analysis*) and with payment (*Information Finding*). This is an interesting outcome, that we hypothesise to be due to the novel work interface, and its relationship with the usual workflow of workers (e.g. in terms of keyboard usage, and cut&paste actions for information finding). This hypothesis will be tested in future work.

Overall, the obtained results are promising. Our takeaway from the whole experimental procedure and our results is that the flexibility (mixed-keyboard input and selection between Web and mobile client) for the interface to be used, the design of the interface, and the task itself are all important factors to consider when building crowdsourcing tasks for conversational interfaces. We believe that the experience with conversational crowd work interfaces could also play a role, but more experiments are needed to understand its relationship with execution time and quality.

We argue that the use of conversational interfaces for crowd work can provide a number of potential benefits, for instance: further democratization of crowd work, as people with limited digital skills or connectivity could then perform retributed digital work [159]; increased workers diversity (in terms of demographics, knowledge, and skills), thus providing better digital experimental environment, e.g. for psychological research [10]; increased workers capacity for low-latency and/or situational microtask crowdsourcing [100, 99, 101, 121]; and push microtask crowdsourcing [143, 19].

### Limitations and Future Work

The recruited workers might not be representative of the whole population of crowd workers. While this risk is mitigated by the popularity of the Figure Eight platform, experiments on other crowdsourcing and messaging platforms are needed for further generalization. To minimize the effect of user interface usability issues, we designed task interfaces that were based on either web or conversational interface. Not all workers were familiar with the Telegram messaging system, but we believe the presence of a web client (identical in functionality and look and feel to the native clients) to have minimized the risk of poor performance due to lack of experience with messaging systems. Issues of task complexity, clarity, and difficulty (tackled, for instance, in [124, 45]) will be addressed in future work. Finally, the experiment included a limited amount of task types and UI elements variations. While we acknowledge such limitations, we believe that our experimental design and results evaluation provide solid answers to the targeted research questions.

## 2.2 Conversational Crowdsourcing Made Easy

The experiments we carried out in Section 2.1 have shown that conversational crowdsourcing could be an effective alternative to the traditional web interface. However the experiments also exposed weaknesses of using a third-party messaging application (i.e., Telegram in Section 2.1) independent of the crowdsourcing platform (i.e. Figure Eight in Section 2.1), which required workers to put extra effort into installing the application and registering a new account in order to complete the task. To this end, in this section, we present TickTalkTurk, a tool that can assist task requesters in quickly deploying crowdsourcing tasks in a customizable conversational worker interface, to further lower the barrier for participation. The conversational worker interface can convey task instructions, deploy microtasks, and gather worker input in a dialogue-based workflow.

In this thesis, TickTalkTurk has been used to carry out crowdsourcing experiments in Section 3.2 *Conversational Styles and Worker Satisfaction*, Section 4.2 *Self-Identification with Worker Avatars*, Section 5.1 *Towards Memorable Information Retrieval*, and Section 5.2 *Understanding Worker Health*. The interface is implemented as a web-based application, which makes it compatible with popular crowdsourcing platforms. The code is available online for the benefit of the community.[11]

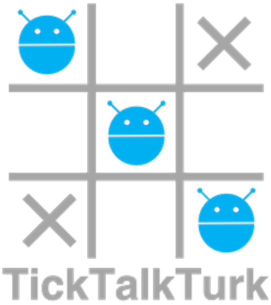
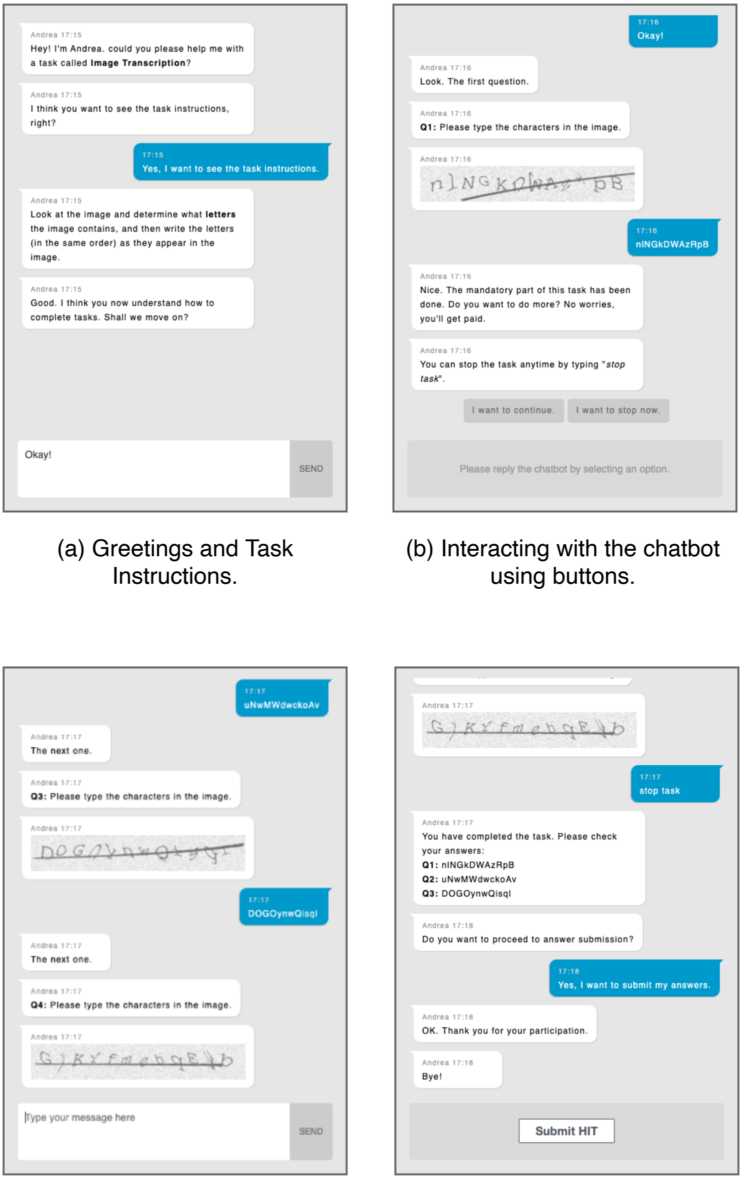


Figure 2.6: The logo of TickTalkTurk.

Advances in microtask crowdsourcing have enabled the possibility of accomplishing complex tasks by relying on crowd workers. Tasks such as image annotation, sentiment analysis, and speech transcription can be easily accomplished on the online crowdsourcing marketplaces. During this process, the crowdsourcing platform is responsible for worker selection, microtask generation, microtask assignment and answer aggregation, while online workers interact with a crowdsourcing system to accept and execute a microtask using a worker interface. A notable feature of the interaction between crowdsourcing platforms and workers in the majority of prior work, is the use of traditional web-based GUIs to communicate with workers, transmit instructions and gather responses thereafter. In the concept of conversational crowdsourcing, based on the conversational agent design in Section 2.1, we improve conversational interfaces compatible with the crowdsourcing platform, further facilitating task execution and task completion.

[11] `https://github.com/qiusihang/ticktalkturk`



Andrea 17:15
Hey! I'm Andrea. could you please help me with a task called **Image Transcription**?

Andrea 17:15
I think you want to see the task instructions, right?

17:15
Yes, I want to see the task instructions.

Andrea 17:15
Look at the image and determine what **letters** the image contains, and then write the letters (in the same order) as they appear in the image.

Andrea 17:15
Good. I think you now understand how to complete tasks. Shall we move on?

Okay!

(a) Greetings and Task Instructions.

17:16
Okay!

Andrea 17:16
Look. The first question.

Andrea 17:16
**Q1:** Please type the characters in the image.

Andrea 17:16

17:16
nINGkDWAzRpB

Andrea 17:16
Nice. The mandatory part of this task has been done. Do you want to do more? No worries, you'll get paid.

Andrea 17:16
You can stop the task anytime by typing "*stop task*".

(b) Interacting with the chatbot using buttons.

17:17
uNwMWdwckoAv

Andrea 17:17
The next one.

Andrea 17:17
**Q3:** Please type the characters in the image.

Andrea 17:17

17:17
DOGOynwQisql

Andrea 17:17
The next one.

Andrea 17:17
**Q4:** Please type the characters in the image.

Andrea 17:17

(c) Interacting with the chatbot using free text.

Andrea 17:17

17:17
stop task

Andrea 17:17
You have completed the task. Please check your answers:
**Q1:** nINGkDWAzRpB
**Q2:** uNwMWdwckoAv
**Q3:** DOGOynwQisql

Andrea 17:18
Do you want to proceed to answer submission?

17:18
Yes, I want to submit my answers.

Andrea 17:18
OK. Thank you for your participation.

Andrea 17:18
Bye!

(d) submitting HIT using a customized HTML component.

Figure 2.7: Two interaction types of the conversational interface.

### Conversational Agent Design

The traditional web-based user interface of a crowdsourcing task typically comprises of two main parts: task instructions and microtasks. Workers are asked to first read instructions and then execute microtasks accordingly. To realize interaction comparable to web-based interfaces, a text-based conversational agent is designed following four main steps: i) task instructions, ii) questions and answers, iii) review, and iv) reward.

***Task instructions.*** Simulating the essence of natural conversation, the conversational agent begins the conversation with greetings, and then presents task instructions (optional), as can be seen in Figure 2.7 (a), via a dialogue with the workers. The goal of this step is to let workers familiarize themselves with the conversational agent and help them understand how to complete the microtasks.

***Questions & Answers.*** The conversational agent asks questions (each question corresponds to a microtask) to workers, and workers can provide responses to microtasks by either typing answers or using customized UI elements (such as buttons).

***Review.*** On the traditional web interface, a worker can easily go back to a question and edit the answer. To realize this affordance in the conversational interface, workers are provided with the opportunity to edit their answers if needed (by typing "edit answer" to enable the answer modification), before submitting the microtasks.

***Reward.*** After reviewing the answers, workers enter the final stage where they can submit their answers and claim their rewards.

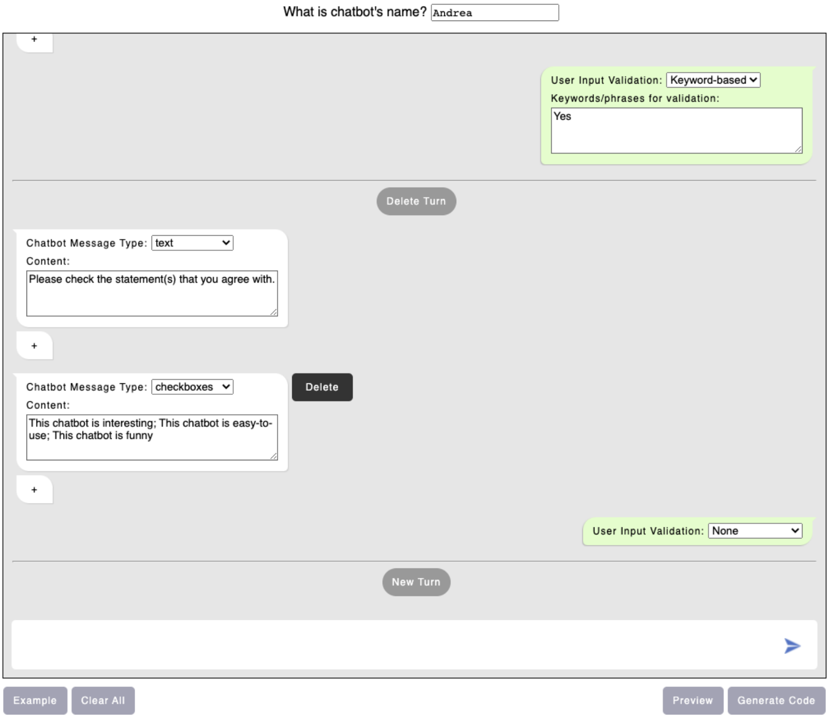
### Web-based Conversational Interface

Popular crowdsourcing platforms (such as Amazon Mechanical Turk and Appen) offer web interfaces based on standard technology like HTML, CSS and Javascript. To avoid the need for installing a messaging application – for instance, Telegram, or Whatsapp, where conversational agents are usually deployed. We designed and implemented the conversational interface in HTML/CSS/Javascript, thus enabling easy integration with existing platforms and access to the available crowd workers.

The conversational interface supports any data source that is supported by HTML5, including text, image, audio, and video. Therefore, most common task types such image classification, sentiment analysis, information finding, object recognition, and speech transcription can all be implemented. Our design provides workers with different means to answer microtasks, as shown in Figure 2.7 (b) and (c). For instance, workers can either type in the textarea or click a button to send their responses. Furthermore, for some tasks that need special functions, UI elements from traditional web pages (e.g. customized buttons, slide bars, drawing tools, etc.) can also be easily ported into conversational interfaces, as shown in Figure 2.7 (d). In addition, the conversational interface can record all the activities of the worker (including all keypress events with timestamps) for further analysis if needed.

### Graphical User Interface for Settings

TickTalkTurk is equipped with a graphical user interface (GUI) for creating a conversational interface for microtask crowdsourcing. The GUI features conversation design and basic customization functions, as shown in Figure 2.8. Users can also make advanced customization, such as changing colors, changing background, and integrating a search engine, using the source code available online (`https://github.com/qiusihang/ticktalkturk`).



What is chatbot's name? Andrea

User Input Validation: Keyword-based

Keywords/phrases for validation: Yes

Chatbot Message Type: text

Content: Please check the statement(s) that you agree with.

Chatbot Message Type: checkboxes

Content: This chatbot is interesting; This chatbot is easy-to-use; This chatbot is funny

User Input Validation: None

Figure 2.8: The setting page of TickTalkTurk.

### Understanding the Components of TickTalkTurk

In the conversation created by TickTalkTurk, the basic component is an *utterance*, which is a message shown in a speech bubble (speech balloon) from either the agent or the user. An *utterance* of the chatbot has four types: text, image, radio buttons, and checkbox.

The agent asks the question(s) using either one *utterance* or multiple *utterances*. The worker gives a response, also through an *utterance*, by using text input, buttons or checkboxes. The question(s) from the agent and the corresponding response from the worker compose a *conversational turn*. A *conversational turn* always starts with the *utterance*(s) of the agent, and ends with an *utterance* of the worker (as a response to the agent).

A conversational crowdsourcing task consists of single/multiple *conversational turn*(s). The agent will proceed to the next *conversational turn* only if the response provided by the worker is validated, using keywords or phrases. If a keyword or a phrase provided is found in the worker's utterance, the conversational agent proceeds to the next turn. Otherwise, the worker will be asked again to provide a valid response.

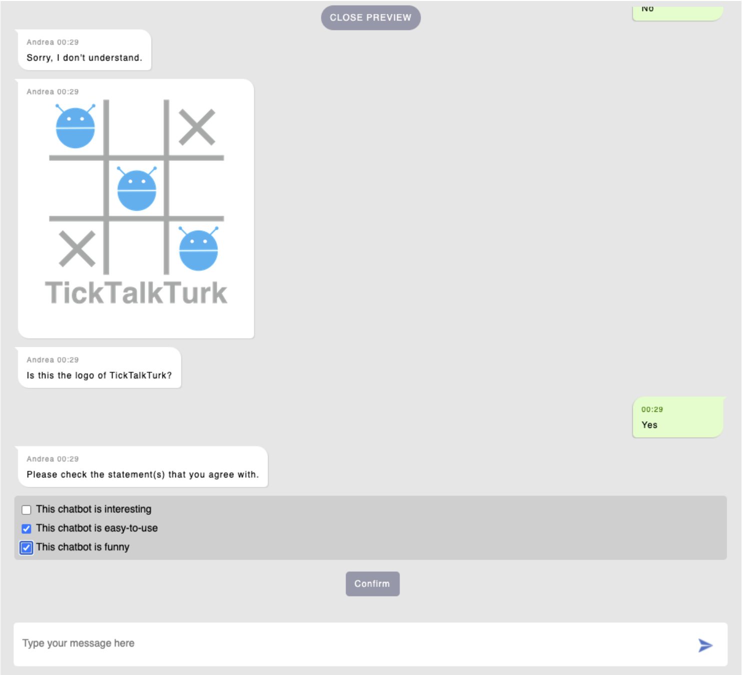
### Creating A Conversational Interface for Crowdsourcing

We present the steps for creating a conversational interface using the GUI of TickTalkTurk.

*1) Customizing an utterance of the agent.* During this step, a task requester can customize the welcome message, task instructions, or microtasks (questions). Four message types are supported, which are text, image, radio buttons, and checkbox. For the text type, requesters can simply input the text that they want to send to the worker in the `message content` field. If requesters would like to send a image, they can input the image URL in the `message content` field. In terms of radio buttons and checkbox, requesters can input all the candidate options in the `message content` field, separated by semicolons (";").

*2) Adding new utterance(s) of the agent.* Task requesters are able to add utterances in a conversational turn. To do this, requesters can click an add (+) button where they want to add an utterance (as shown in Figure 2.8), and then repeat the procedure described in Step 1. Requesters can also delete an utterance by clicking the corresponding `Delete` button. Please note that the first utterance of a conversational turn cannot be deleted.

*3) Adding validation for worker input.* In order to proceed to the next conversational turn, for instance, moving to the next microtask after completing one microtask (or after reading the task instructions), the worker input should be validated to ensure that the agent is ready to move on. Requesters can choose either none validation or keyword-based validation. None validation is supposed to be used for free-text input, it allows any string input except strings only containing spaces. Keyword-based validation will check whether the worker input contains the pre-defined keyword or phrases provided by requesters.



No

Andrea 00:29
Sorry, I don't understand.

Andrea 00:29

Andrea 00:29
Is this the logo of TickTalkTurk?

00:29
Yes

Andrea 00:29
Please check the statement(s) that you agree with.

□ This chatbot is interesting
☑ This chatbot is easy-to-use
☑ This chatbot is funny

Figure 2.9: The preview function (for testing the customized chatbot) of TickTalkTurk.

*4) Adding new conversational turn(s).* As shown in Figure 2.8, requesters can click a `New Turn` button where they want to add a conversational turn. Afterward, requesters will repeat Step 1 - Step 3 to customize the new conversational turn. Requesters can also delete the new conversational turn by clicking the corresponding `Delete Turn` button.

*5) Conversational crowdsourcing task preview.* After adding all the necessary conversational turns, which may represent task instructions, microtasks (questions), goodbye messages, etc., task requesters can preview the conversational crowdsourcing task in the GUI of TickTalkTurk by clicking the `Preview` button, as shown in Figure 2.9. The preview function only works if there is no error. Otherwise, all the errors will be highlighted in red in the GUI.

*6) Publishing the conversational crowdsourcing task.* Task requesters can use the `Generate Code` button to download the HTML source code of the conversational interface. To publish tasks on Amazon's Mechanical Turk or Toloka, requesters can simply paste the HTML code and launch the task on the platforms. To publish tasks on Prolific, requesters can host the web page generated by TickTalkTurk on their own servers (or other web services, e.g., Heroku[12] and Netlify[13]), and use its URL link on the platform to redirect workers to the conversational interface. This method also applies to Mechanical Turk and Toloka.

### Design Implications

The conversational interface generated by TickTalkTurk is purely HTML based. Elements used in traditional web interfaces can be easily ported into conversational interfaces. With TickTalkTurk, the overheads of designing and implementing conversational interfaces can be easily reduced. Task requesters can quickly deploy and publish their tasks on popular crowdsourcing platforms, to obtain not only high-quality outcomes but also an increase in worker engagement and a better understanding of worker performance (which will be further explained in the next chapters).

### Limitations and Future Work

TickTalkTurk currently does not support complex conversation design. Future work could focus on the crowdsourcing conversation design with novel features, such as branching dialogues, switching conversational styles, and understanding natural language. Furthermore, although the current version of TickTalkTurk can be used to execute common crowdsourcing microtasks, the conversational interface is yet to be completely compatible with novel task design such as crowd-mapping [176]. In the imminent future, we will further update TickTalkTurk to make it more programmable and extensible.

[12] `https://www.heroku.com/`

[13] `https://www.netlify.com/`

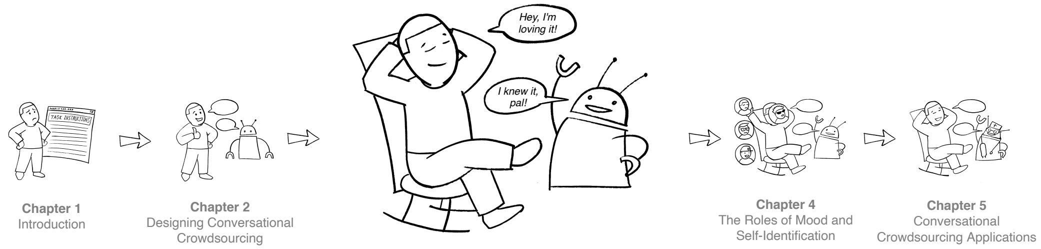
## 2.3 Chapter Summary

In this chapter, we conducted a systematic analysis across five task types to investigate the utility of conversational crowdsourcing. We showed that task execution times and output qualities of conversational crowdsourcing are comparable to the ones achievable through web based interfaces. The workers recruited in our experiments expressed positive opinions towards this work execution medium. To further lower the barrier for participation and make the conversational interface more compatible with prevailing crowdsourcing platforms, we designed TickTalkTurk to help task requesters conveniently deploy conversational crowdsourcing tasks on prevailing crowd work marketplaces. The conversational interface used in TickTalkTurk is purely HTML based, elements used in traditional web interfaces can be easily ported into conversational interfaces. With TickTalkTurk, the overheads of designing and implementing conversational interfaces can be easily reduced.

We provide evidences of the suitability of conversational crowdsourcing in current crowd work marketplaces. We highlight the importance of task-specific interaction design, but also the convenience of advanced text input interfaces currently available in conversational interfaces. The continuous evolution of the functionalities available in such platforms (e.g. novel content types, micropayment, etc.) could allow a broader, more democratic, and potentially decentralised adoption of crowd work (both for offer and demand).

# Chapter 3

# Engagement and Satisfaction


TASK INSTRUCTIONS
Hey, I'm loving it!
I knew it, pal!
Chapter 1
Introduction
Chapter 2
Designing Conversational Crowdsourcing
Chapter 4
The Roles of Mood and Self-Identification
Chapter 5
Conversational Crowdsourcing Applications

Crowdsourcing marketplaces have provided a large number of opportunities for online workers to earn a living. To improve satisfaction and engagement of such workers, who are vital for the sustainability of the marketplaces, in Chapter 2, we have shown that conversational interfaces could be an alternative to the traditional web interface for task execution. The rationale behind using conversational interfaces stems from the potential engagement that conversation can stimulate. However, the impact of conversational crowdsourcing on worker satisfaction and engagement remains unexplored. Furthermore, we noticed that prior works in linguistics have shown that '*conversational styles*' can play an important role in communication. There are unexplored opportunities to study conversational styles in conversational crowdsourcing with an end goal of improving worker satisfaction, engagement, and quality. Therefore, we drew from Deborah Tannen's theory of conversational style, which classifies the style broadly into *High-Involvement* and *High-Considerateness*. A *High-Involvement* style refers to enthusiasm, fast pace, personal topics, and so on, while a *High-Considerateness* style refers to respect, hesitation, longer pauses, and so on. In this chapter, we performed two studies about how conversational crowdsourcing and conversational styles could affect worker engagement and satisfaction.

In the first study (Section 3.1), we investigated the effectiveness of using conversational crowdsourcing to improve worker engagement. 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 could be affected by the conversational agent and its conversational styles. Our results revealed that conversational interfaces could be effective in engaging workers, and a suitable conversational style had the potential to improve worker engagement.

In the second study (Section 3.2), we investigated the role of workers' conversational styles in conversational microtask crowdsourcing. To this end, we used TickTalkTurk to support task execution, and we proposed methods to estimate the conversational style of a worker. Our experimental setup was designed to empirically observe how conversational styles of workers related with quality-related outcomes. Results showed that even a naive supervised classifier could predict the conversation style with high accuracy, and crowd workers with an Involvement conversational style provided a significantly higher output quality, exhibited a higher user engagement and perceived less cognitive task load in comparison to their counterparts.

In this chapter, both studies showed that conversational crowdsourcing had positive effects on the worker satisfaction and engagement, and revealed the importance of applying suitable conversational styles. Our findings have important implications on workflows and task design with respect to improving worker performance and their engagement in microtask crowdsourcing.

The content of this chapter is based on the following papers:

Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon. Improving Worker Engagement Through Conversational Microtask Crowdsourcing. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1-12, 2020. (Section 3.1 is based on this paper)

Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon. Estimating Conversational Styles in Conversational Microtask Crowdsourcing. Proceedings of the ACM on Human-Computer Interaction (CSCW), vol. 4, pp. 1-23, 2020. (Section 3.2 is based on this paper)

## 3.1 Improving Worker Engagement

In Chapter 2, we have explored the suitability of conversational interfaces for microtask crowdsourcing by juxtaposing them with standard web interfaces in a variety of popularly crowdsourced tasks. We 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 crowdsourcing on the engagement of workers. Previous works have studied the nature of tasks that are popularly crowdsourced on MTurk, showing that tasks are often deployed in large batches consisting of similar HITs (human intelligence tasks) [3, 51]. Crowdsourcing microtasks can often be monotonous and repetitive in nature. 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 [186, 42, 61]. Although researchers have already paid attention to worker engagement and retention, there is a lack of understanding of whether conversational crowdsourcing would either alleviate or amplify the concerns surrounding worker engagement. In this section, we aim to fill this knowledge gap.

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

**RQ3.1:** To what extent can conversational crowdsourcing improve the worker engagement?

**RQ3.2:**How do different conversational styles affect the performance of workers and their cognitive load in conversational crowdsourcing?

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

We first investigated the effect of conversational interfaces with different conversational styles on quality related outcomes in comparison to the traditional web interfaces. We addressed **RQ3.1** 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 [164, 244]. We addressed **RQ3.2** 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 showed that conversational crowdsourcing had 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 had the potential to engage workers further (in specific task types), although our results were inconclusive in this regard. This work takes crucial strides towards furthering the understanding of conversational crowdsourcing, revealing insights into the role of conversational styles across a variety of tasks.

### Method: Conversational Interface for Crowdsourcing

We designed and implemented conversational interfaces that enabled 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.[14]

#### Workflow of Conversational Microtasking

The conversational interface was designed based on the workflow proposed in Chapter 2, to help workers in carrying out crowdsourcing tasks. The main building blocks of conversational microtasks are similar to Figure 2.2, including 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 3.1 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 MTurk, the worker can claim the corresponding monetary compensation, proportional to the number of answered questions.

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

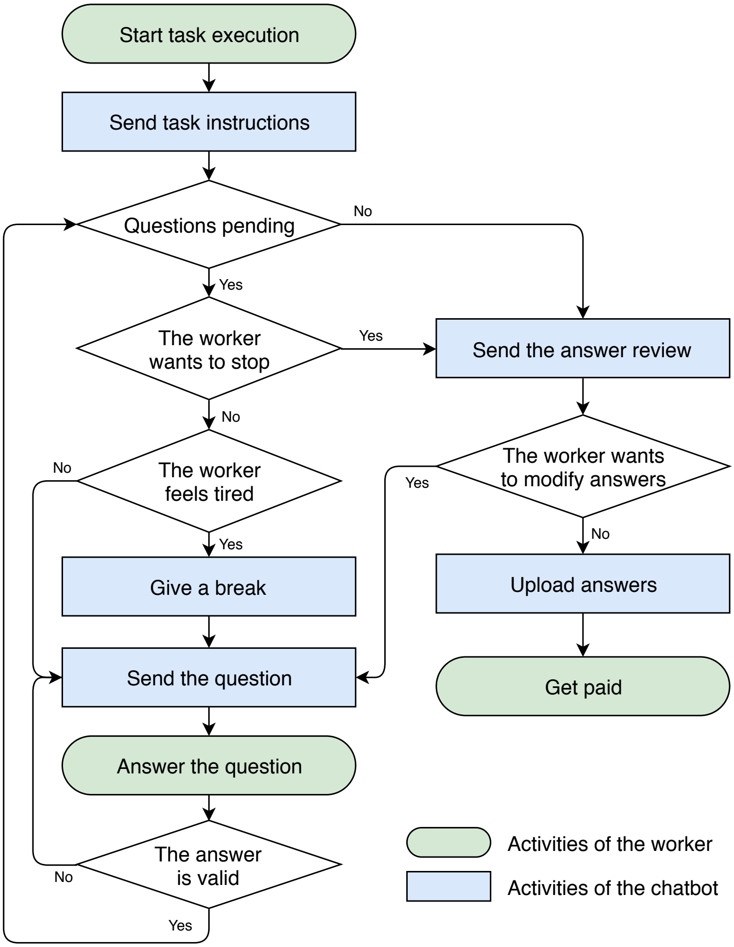


Figure 3.1: The workflow of conversational microtask crowdsourcing.

### Conversational Styles: Involvement or Considerateness

We used Deborah Tannen's theory of conversational styles in our study [210, 211]. Tannen's analysis of conversational style was 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 [195]. More-

Table 3.1: 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 |

over, 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 [95].

Tannen identified four main features of the conversational style, namely *topic*, *pacing*, *narrative strategies*, and *expressive paralinguistics* [211]. 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 3.1. The criteria can be organised into two categories:

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

**2) Content (C3, C4, C5, C6):** The conversational agent corresponding to the High-Involvement style introduces a new topic to the worker (for instance, telling workers how to answer questions, how to edit answers, and how to submit answers) without hesitation (C3). On the contrary, we use some words or paralanguage such as "*Well..*" and "*Hmm..*" to simulate the hesitation of the High-Considerateness conversational agent (C3). Furthermore, the conversational agent of High-Involvement style uses less syntax (C4) and chats directly (C5), while the agent of High-Considerateness style uses relatively complex syntax (C4) and tends to express ideas/topics in an indirect or polite way (C5). Tannen also emphasized the importance of asking questions for the High-Involvement style [211]. 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 3.2.

### 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 [195, 215]. 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.

Table 3.2: 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 3.1).

| ***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 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 3.2). 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 3.2), 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 [62, 206]. Note that *actual options* are invisible to workers. The only *actual option* corresponding to "opening greetings" is `go ahead`, while the only *actual option* of "time requirement" is `understand the time requirement`. For the "task instructions" interaction, there are two *actual options*: `show instructions` and `skip instructions`, where the former elucidates how to answer the crowdsourcing question and the latter directly leads the worker through to the task execution stage. **2)** As *actual options* are invisible to workers, we create two *visible options* (referring to High-Involvement and High-Considerateness respectively) for

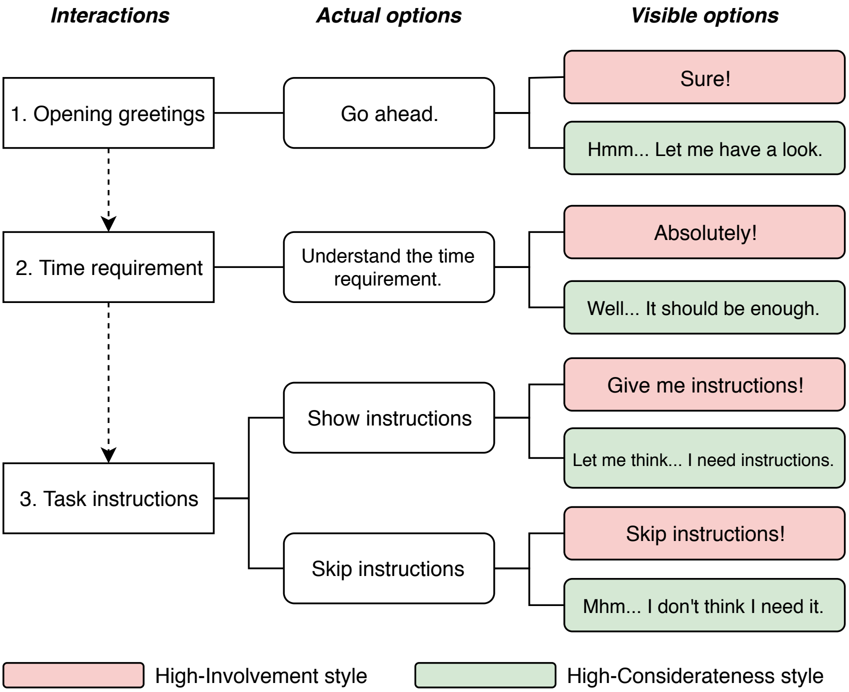


Figure 3.2: Options given to the worker for conversational style estimation and alignment.

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 MTurk. This will allow us to analyze our results in the light of the findings from Chapter 2. 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 3.3): Information Finding, Sentiment Analysis, CAPTCHA Recognition, and Image Classification [72].

Table 3.3: Summary of task types.

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

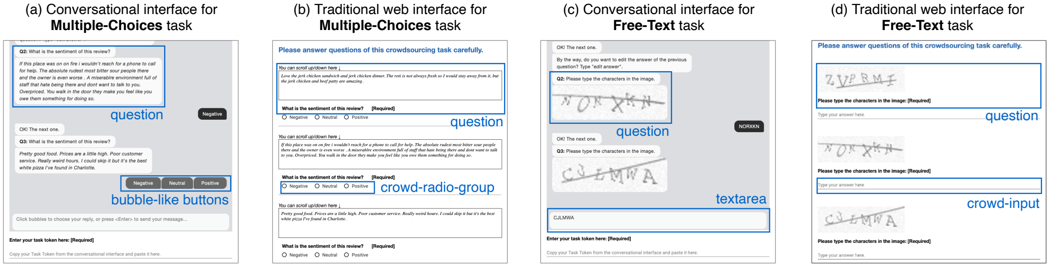
*Information Finding (IF).* Workers are asked to find a given store on Google Maps and report its rating (i.e., the number of stars). The information corresponding to stores is obtained from a publicly available Yelp dataset[15].

*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[16], 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, panda, pigeon, and rooster) selected from Caltech101 Dataset [60]. They are tasked with determining which animal a given image contains, and selecting the corresponding option.

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



(a) Conversational interface for **Multiple-Choices** task

**Q2:** What is the sentiment of this review?

*If this place was on on fire i wouldn't reach for a phone to call for help. The absolute rudest most bitter sour people there and the owner is even worse . A miserablre environment full of staff that hate being there and dont want to talk to you. Overpriced. You walk in the door they make you feel like you owe them something for doing so.*

Negative

OK! The next one.

**Q3:** What is the sentiment of this review?

*Pretty good food. Prices are a little high. Poor customer service. Really weird hours. I could skip it but it's the best white pizza I've found in Charlotte.*

Click bubbles to choose your reply, or press <Enter> to send your message...

**Enter your task token here: [Required]**

Copy your Task Token from the conversational interface and paste it here.

(b) Traditional web interface for **Multiple-Choices** task

**Please answer questions of this crowdsourcing task carefully.**

*You can scroll up/down here ↓*

*Love the jerk chicken sandwich and jerk chicken dinner. The roti is not always fresh so I would stay away from it, but the jerk chicken and beef patty are amazing.*

**What is the sentiment of this review? [Required]**

○ Negative ○ Neutral ○ Positive

*You can scroll up/down here ↓*

*If this place was on on fire i wouldn't reach for a phone to call for help. The absolute rudest most bitter sour people there and the owner is even worse . A miserablre environment full of staff that hate being there and dont want to talk to you. Overpriced. You walk in the door they make you feel like you owe them something for doing so.*

**What is the sentiment of this review? [Required]**

○ Negative ○ Neutral ○ Positive

*You can scroll up/down here ↓*

*Pretty good food. Prices are a little high. Poor customer service. Really weird hours. I could skip it but it's the best white pizza I've found in Charlotte.*

**What is the sentiment of this review? [Required]**

○ Negative ○ Neutral ○ Positive

(c) Conversational interface for **Free-Text** task

OK! The next one.

By the way, do you want to edit the answer of the previous question? Type "*edit answer*".

**Q2:** Please type the characters in the image.

NORXKN

OK! The next one.

**Q3:** Please type the characters in the image.

CJLMWA

**Enter your task token here: [Required]**

Copy your Task Token from the conversational interface and paste it here.

(d) Traditional web interface for **Free-Text** task

**Please answer questions of this crowdsourcing task carefully.**

**Please type the characters in the image: [Required]**

Type your answer here.

**Please type the characters in the image: [Required]**

Type your answer here.

**Please type the characters in the image: [Required]**

Type your answer here.

Figure 3.3: The comparison of conversational interfaces embedded on the user interface of MTurk and traditional web interfaces using HTML elements provided by MTurk, where the worker needs to provide a Task Token acquired from the conversational interface after the task is completed.

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

[16] `https://github.com/kuszaj/claptcha`

### 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 [42, 186], 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) [163, 164]. The UES-SF contains four sub-scales with 12 items, which is a tool which is a widely used tool for measuring user engagement for measuring user engagement in HCI contexts. Each item is measured by a 7-pt Likert-scale from "*1: Strongly Disagree*" to "*7: Strongly Agree*". 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[17]. 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 [89].

### Worker Interface

Both the web interface and the conversational interface are designed and implemented on top of MTurk (see Figure 3.3). For both interfaces, the demographic survey, UES-SF and NASA-TLX are created using default HTML-based questions widgets provided by MTurk; using the *Crowd HTML Element*.

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

We designed and implemented the conversational agent purely based on HTML and Javascript. Thus, it can be perfectly embedded on the MTurk task page without any re-

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

strictions. 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 3.3, we developed a rule-based conversational agent based on chat-bubble[18].

### 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$. The reward for the optional questions is given to workers through the "bonusing" function on MTurk. We estimated the execution time and paid workers 0.01$ 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 bonuses at the onset, ensuring that there was no unnatural financial uncertainty other than what is typical on MTurk.

### 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 by Holm-Bonferroni method [98]).

*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

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

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 MTurk as their primary source of income, while about half of the workers (55.8%) reported that MTurk 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 3.2. The number of workers whose conversational styles were estimated as High Involvement and High Considerateness are shown in Figure 3.4.

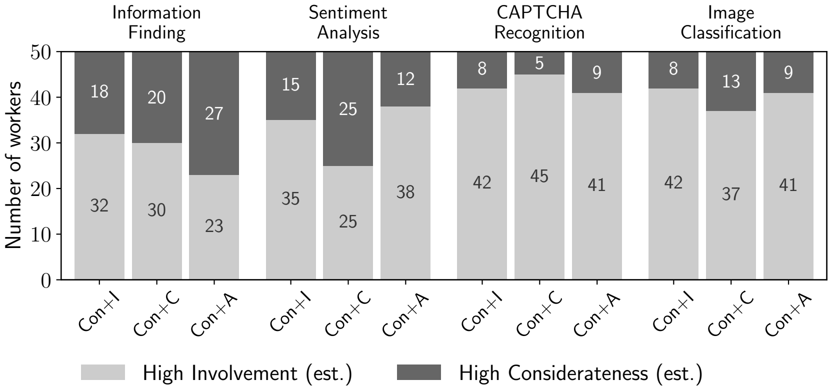


Figure 3.4: Distributions of estimated styles across all conditions.

As we described earlier, the conversational agent maintained a High-Involvement and High-Considerateness styles in Con+I and Con+C conditions respectively, while in Con+A conditions the conversational agent initiated with either Involvement or Considerateness style randomly. Figure 3.5 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).

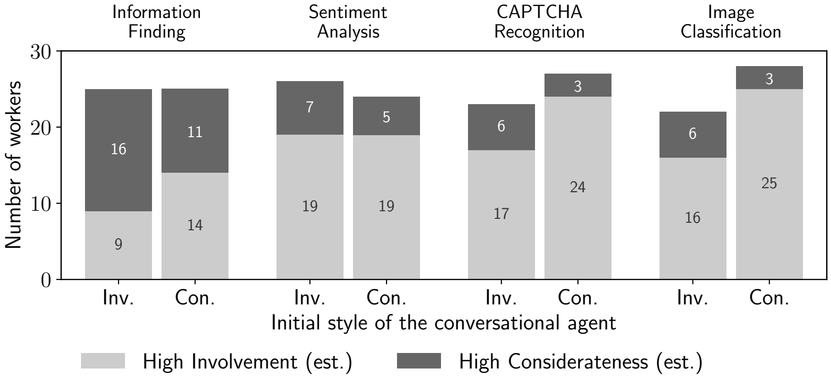


Figure 3.5: Distributions of estimated styles of conversational interfaces with style alignment by two initial styles.

### 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 3.4 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 Chapter 2. 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 3.4: Worker accuracy ($\mu \pm \sigma$: mean and standard deviation) and $p$-values across different task types and interface conditions.

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

### 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.

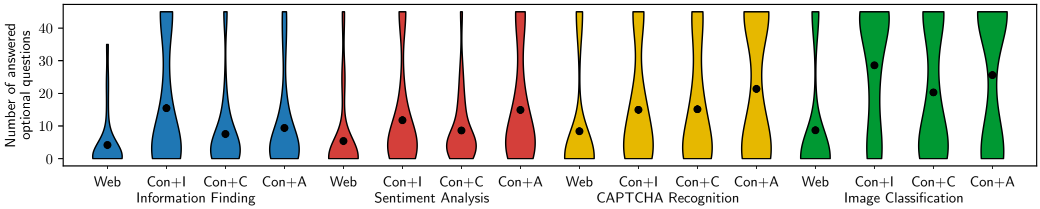


Figure 3.6: A violin plot representing the number of optional questions answered by workers across different task types and different interfaces, where the black dots represent the mean value. A violin plot is a hybrid of a box plot and a kernel density plot, revealing peaks in the data that cannot be visualized using box plots.

Figure 3.6 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 3.5. 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 (RQ3.2). 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 (RQ3.2). 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 (RQ3.1).

Table 3.6 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.

***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 3.7 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 (RQ3.1) with different conversational styles (RQ3.2) show that the UES-SF scores have no significant difference across four interfaces (conversational styles) within each task type.

Table 3.5: 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**<br>(vs. Con+I,C,A) | **Con+I**<br>(vs. Con+C,A) | **Con+C**<br>(vs. Con+A) | **Con+A** |
|---|---|---|---|---|
| *IF* | $4.18 \pm 8.66$<br>($p =$ **1.8e-4***, 0.02, 5.1e-3) | $15.47 \pm 18.0$<br>($p = 0.08, 0.17$) | $7.55 \pm 12.14$<br>($p = 0.67$) | $9.4 \pm 13.63$ |
| *SA* | $5.4 \pm 11.99$<br>($p =$ **2.7e-5***, **8.3e-5***, **2.3e-6***) | $11.78 \pm 15.26$<br>($p = 0.6, 0.29$) | $8.63 \pm 10.32$<br>($p = 0.09$) | $14.92 \pm 16.09$ |
| *CR* | $8.4 \pm 16.75$<br>($p =$ **1.3e-3***, 2.3e-3, **2.0e-4***) | $14.96 \pm 16.73$<br>($p = 0.98, 0.22$) | $15.14 \pm 16.8$<br>($p = 0.19$) | $21.37 \pm 19.9$ |
| *IC* | $8.7 \pm 17.17$<br>($p =$ **1.2e-6***, **6.1e-5***, **2.9e-5***) | $28.6 \pm 17.97$<br>($p = 0.07, 0.67$) | $20.29 \pm 19.08$<br>($p = 0.34$) | $25.61 \pm 20.52$ |

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

Table 3.6: 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%) |

Table 3.7: 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* | $4.12 \pm 1.40$ | $4.39 \pm 1.44$ | $3.68 \pm 1.45$ | $3.81 \pm 1.53$ | $\mathbf{3.98 \pm 1.51}$ | $4.12 \pm 1.30$ | $4.43 \pm 1.20$ | $4.07 \pm 1.46$ | $4.28 \pm 1.38$ | $\mathbf{4.21 \pm 1.37}$ |
| *Perceived usability* | $3.71 \pm 1.67$ | $3.70 \pm 1.61$ | $3.86 \pm 1.70$ | $4.24 \pm 1.61$ | $\mathbf{3.86 \pm 1.68}$ | $3.91 \pm 1.83$ | $3.86 \pm 1.67$ | $4.19 \pm 1.63$ | $4.42 \pm 1.85$ | $\mathbf{4.08 \pm 1.78}$ |
| *Aesthetic appeal* | $4.23 \pm 1.46$ | $4.29 \pm 1.29$ | $4.10 \pm 1.12$ | $4.01 \pm 1.58$ | $\mathbf{4.14 \pm 1.40}$ | $4.75 \pm 1.28$ | $4.67 \pm 1.51$ | $4.84 \pm 1.31$ | $4.86 \pm 1.12$ | $\mathbf{4.76 \pm 1.35}$ |
| *Reward factor* | $4.35 \pm 1.23$ | $4.44 \pm 1.53$ | $4.41 \pm 1.33$ | $4.17 \pm 1.49$ | $\mathbf{4.33 \pm 1.44}$ | $4.99 \pm 1.23$ | $4.90 \pm 1.33$ | $4.95 \pm 1.31$ | $5.05 \pm 1.37$ | $\mathbf{4.95 \pm 1.36}$ |
| *Overall* | $4.10 \pm 0.85$ | $4.21 \pm 0.85$ | $4.01 \pm 0.69$ | $4.06 \pm 1.00$ | $\mathbf{4.07 \pm 0.90}$ | $4.44 \pm 0.87$ | $4.46 \pm 0.98$ | $4.51 \pm 0.90$ | $4.65 \pm 0.88$ | $\mathbf{4.50 \pm 0.97}$ |
| | **CAPTCHA Recognition** | | | | | **Image Classification** | | | | |
| *Focused attention* | $3.83 \pm 1.76$ | $3.92 \pm 1.61$ | $3.93 \pm 1.56$ | $4.39 \pm 1.55$ | $\mathbf{4.00 \pm 1.66}$ | $4.30 \pm 1.45$ | $4.21 \pm 1.77$ | $4.35 \pm 1.62$ | $4.16 \pm 1.85$ | $\mathbf{4.23 \pm 1.70}$ |
| *Perceived usability* | $4.95 \pm 1.57$ | $4.71 \pm 1.66$ | $4.56 \pm 1.64$ | $4.74 \pm 1.43$ | $\mathbf{4.71 \pm 1.62}$ | $4.41 \pm 1.93$ | $4.91 \pm 1.53$ | $4.90 \pm 1.67$ | $4.68 \pm 1.78$ | $\mathbf{4.70 \pm 1.77}$ |
| *Aesthetic appeal* | $3.74 \pm 1.71$ | $4.10 \pm 1.73$ | $3.95 \pm 1.81$ | $3.94 \pm 1.69$ | $\mathbf{3.92 \pm 1.76}$ | $4.73 \pm 1.42$ | $4.53 \pm 1.56$ | $4.75 \pm 1.37$ | $4.75 \pm 1.65$ | $\mathbf{4.67 \pm 1.54}$ |
| *Reward factor* | $4.43 \pm 1.71$ | $4.42 \pm 1.79$ | $4.50 \pm 1.68$ | $4.25 \pm 1.66$ | $\mathbf{4.38 \pm 1.74}$ | $4.97 \pm 1.32$ | $5.09 \pm 1.73$ | $4.87 \pm 1.56$ | $5.14 \pm 1.60$ | $\mathbf{5.00 \pm 1.60}$ |
| *Overall* | $4.24 \pm 1.27$ | $4.29 \pm 1.11$ | $4.23 \pm 1.12$ | $4.33 \pm 1.14$ | $\mathbf{4.25 \pm 1.20}$ | $4.60 \pm 0.91$ | $4.69 \pm 1.20$ | $4.72 \pm 1.03$ | $4.68 \pm 1.19$ | $\mathbf{4.65 \pm 1.13}$ |

However, as shown in Table 3.8 ($p$-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.

Table 3.8: $p$-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)

***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 RQ3.2, we calculated and listed unweighted NASA-TLX scores in Table 3.9. According to pairwise independent t-tests (expected $\alpha = 0.05$, two-tailed, corrected by Holm-Bonferroni method), the NASA-TLX scores have no significant difference across four interfaces (conversational styles) within each task type. However the conversational interface with aligned style has the potential to reduce the cognitive task load for *Information Finding* task compared with the web interface (no significance, $p = 0.033$, which is less than 0.05 but higher than corrected $\alpha$).

Table 3.9: 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**<br>(vs. Con+I,C,A) | **Con+I**<br>(vs. Con+C,A) | **Con+C**<br>(vs. Con+A) | **Con+A** |
|---|---|---|---|---|
| *IF* | $52.35 \pm 20.75$<br>($p = 0.51$, 0.12, 0.03) | $49.62 \pm 20.39$<br>($p = 0.37$, 0.13) | $46.05 \pm 19.63$<br>($p = 0.51$) | $43.4 \pm 20.25$ |
| *SA* | $50.27 \pm 17.76$<br>($p = 0.95$, 0.31, 0.17) | $50.02 \pm 20.54$<br>($p = 0.37$, 0.22) | $46.54 \pm 18.26$<br>($p = 0.71$) | $45.15 \pm 18.85$ |
| *CR* | $38.23 \pm 19.56$<br>($p = 0.81$, 0.74, 0.6) | $37.29 \pm 20.26$<br>($p = 0.58$, 0.78) | $39.54 \pm 20.2$<br>($p = 0.4$) | $36.14 \pm 19.89$ |
| *IC* | $43.38 \pm 22.64$<br>($p = 0.46$, 0.07, 0.44) | $40.22 \pm 19.56$<br>($p = 0.23$, 0.94) | $35.57 \pm 19.11$<br>($p = 0.3$) | $39.89 \pm 21.94$ |

### Discussion

Aspects such as task complexity [236], task types, instructions [75] are instrumental in shaping crowd work [117]. However, previous work has shown that conversational interfaces can effectively benefit workers from different perspectives [128, 99]. 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 tended 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 (RQ3.2). The results of style alignment further showed that workers' conversational styles were 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 [236] in comparison to *CAPTCHA Recognition* and *Image Classification.* This calls for further exploration of the impact of task complexity on task outcomes within conversational 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 suggested that a High-Involvement conversation style could provide workers with engagement stimuli for long-term retention (RQ3.2).

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 suggested that the input type (free text vs. multiple choices) had a principal impact on the UES-SF scores, which weaken the effect of different interfaces and conversational styles.

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 suggested that the task complexity had an impact on the perceived cognitive load.

### Design Implications

We found that workers tended 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 suggested that giving the conversational agent a High-Involvement style could generally improve the worker retention in conversational 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.

### Limitations and Future Work

Our findings with respect to the impact of conversational interfaces on worker engagement across different task types suggested that different conversational styles of the agent could affect the worker retention, albeit not consistently. Moreover, further experiments that decouple the impact of task difficulty [236] 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 [50]. 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 [86].

*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 crowdsourcing 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.

## 3.2 Conversational Styles and Worker Satisfaction

Previous works in the field of linguistics and psychology have shown the important role that *conversational styles* have on inter-human communication [125, 210, 211]. Having been developed in the context of human conversations, the insights and conclusions of these works are not directly applicable to conversational crowdsourcing, since the contrasting goal of workers is to optimally allocate their effort rather than being immersed in conversations. Previous work studied how specific linguistic strategies and language styles can affect work outcomes [193, 113]. In Section 3.1, we investigated whether different conversational styles of an agent can increase worker engagement. However, to the best of our knowledge, current conversational agents (particularly for crowdsourcing) have only studied the effects of the conversational style of agents, rather than the conversational style of online users (i.e., workers in the context of microtask crowdsourcing). Understanding the role of workers' conversational styles in human computation can help us better adapt strategies to improve output quality and worker engagement, or better assist and guide workers in the training process. To this end, there is the need for novel methods for the classification of conversational styles in the context of microtask crowdsourcing. In this section, we delve into the following research questions:

**RQ3.3:** How can the conversational style of a crowd worker be reliably estimated?

**RQ3.4:** To what extent does the conversational style of crowd workers relate to their work outcomes, perceived engagement, and cognitive task load in different types of tasks?

We used TickTalkTurk (Section 2.2) to support crowdsourcing task execution and extraction of linguistic features from the text-based conversation between the user and the agent. We designed a coding scheme according to Tannen's theory [210, 211] and corresponding to conversational styles based on the five dimensions of linguistic devices that have been examined. Demonstrating the practical utility of our findings in this study, we proposed methods to predict the conversational style of users using either rich conversational data, or limited conversational data.

To evaluate our methods, we recruited 180 unique online crowd workers from Amazon Mechanical Turk (MTurk) and conducted experiments to investigate the feasibility of conversational style estimation for online crowdsourcing. We also analyzed the impact of conversational style on output quality, worker engagement (using worker retention and the User Engagement Scale), and perceived task load (using the NASA-TLX instrument). Our results showed that we could predict the conversational style of workers using rich conversation data with a high accuracy (gradient boost: 80%), while we could also predict their conversational style using limited conversation data with an acceptable accuracy (gradient boost: 67%). Furthermore, our experimental findings revealed that workers with an *Involvement* conversational style had significantly higher output quality, higher user engagement and less cognitive task load while they were completing a high-difficulty task, and had less task execution time in general. The findings have important implications on worker performance prediction, task scheduling and assignment in microtask crowdsourcing. To the best of our knowledge, this is the first work that explores the impact of conversational style on quality-related outcomes in conversational microtasking, and proposes methods to estimate the conversational style of users. To facilitate further research and for the benefit of the CSCW and HCI community, we publicly released our data and code[19].

### Conversational Style Estimation

Emulating particular conversational styles suitable to given contexts, or aligning the conversational style of an agent to the preferred style of workers, may help to improve worker engagement, satisfaction, and even output quality. To enable further research in this direction, we first need a reliable method to estimate the conversational style of a worker. Therefore, we first introduce background work on conversational styles, and present a novel coding scheme designed to label and estimate conversation style of workers in conversational microtasking. We then propose two approaches for conversational style prediction, based on rich and limited conversation data from workers.

#### High Involvement and High Considerateness

We used Tannen's theory for classifying conversational styles [210, 211]. According to her theory, conversational styles emerge through the combined use of different linguistic devices. At the end of her book, Tannen identifies nine *dimensions* of linguistic devices that are related to conversational styles: *Personal focus of topic*, *Paralinguistic features*, *Enthusiasm*, *Use of questions*, *Pacing*, *Use of repetition*, *Topic cohesion*, *Tolerance of silence*, and *Laughter* [211]. She then presented an example of how conversational styles could be estimated: she created four continua (which could be extended, if needed) corresponding to four linguistic devices mapped on the nine dimensions above: humor, pace, percentage of narrative turns, and narratives [211]. These continua are superimposed upon one another to get an overall continuum. Participants who receive high scores on the overall continuum are classified as High-Involvement speakers, while those who receive low scores are classified High-Considerateness speakers.

#### Coding Scheme of Conversational Style

While providing a conceptual framework for the definition and characterisation of conversational styles, Tannen's theory is not directly applicable to conversational crowdsourcing. Tannen's work was developed (and tested) in the context of human conversations, which are typically long and articulated. In conversational crowdsourcing , devices like "humor" and "the percentage of narrative turns" are clearly at odds with the need for workers to optimally allocate their effort. Moreover, Tannen's continua-based method for conversational style estimation does not have specific criteria to guide readers to distribute speakers on continua. For these reasons, a novel coding scheme for systematically classifying the conversational style is required, to enable the classification of coding styles, and guide the creation of ground truth data for conversation style estimation. This coding scheme builds upon a subset of the linguistic dimensions listed in the previous section. We exclude *Paralinguistic features*, *Use of repetition* and *Laughter*.

Several paralinguistic features, such as pitch shifting and voice quality, are usually absent in text-based chat. Repetition is fairly often used in voice chats, but it is absent in text chats [107], which can be explained by the ability for users to see previous utterances on the computer screen. Finally, we ignore the effects of text-based laughter as previous work

[19] `https://qiusihang.github.io/convsty`

suggests that nonlinguistic (nonverbal) cues (such as smile, laughter) are mostly absent from text-based conversations [188].

Table 3.10: Coding scheme for conversational style.

| ***Dimension*** | **Score** | **Criteria** |
|---|---|---|
| *1) Personal focus of topic* | 1 | The worker prefers responding to the questions with personal opinions or personal anecdotes. For example, the worker uses first-person pronouns and phrases such as “I think”, “I like”, “my experience”. |
| | -1 | The worker prefers responding to questions by using objective descriptions. For example, using impersonal phrases such as “it is”. |
| *2) Enthusiasm* | 1 | The worker demonstrates a willingness to converse with the conversational agent. For example, by responding positively to questions from the agent that would prolong the conversation. |
| | -1 | The worker appears to be disinterested in the conversation with the agent. For example, by constantly seeking to end the conversation and responding with “no more”, “nothing else”, or similar phrases. |
| *3) Pacing* | 1 | Calculate the mean *pace* (typing rate) of all the workers. The score of the worker whose mean *pace* $\geq$ *median* is 1 (relatively faster pace). |
| | -1 | Calculate the mean *pace* of all the workers. The score of the worker whose mean *pace* $<$ *median* is -1 (relatively slower pace). |
| *4) Tolerance of silence* | 1 | Calculate the mean *percentage of self-editing* (fractions of deleted characters among all the typed characters) of all the workers. The score of the worker whose mean *percentage of self-editing* $<$ *median* is 1. |
| | -1 | Calculate the mean *percentage of self-editing* of all the workers. The score of the worker whose mean *percentage of self-editing* $\geq$ *median* is -1. |
| *5) Topic cohesion & Use of questions* | 1 | The worker prefers to express opinions directly linked to the topic or asks questions when in doubt. |
| | -1 | The worker deviates from the topic without asking questions, but by responding respectfully to the conversational agent when in doubt. |

We include *Tolerance of silence* in the coding scheme, i.e. hesitation and silence occurring in conversations, but with some adaptation. In text-based chat, we measure tolerance of silence through editing actions (i.e., when users edit a message before it is sent). We calculate the percentage of deleted keys among all the keys pressed by the worker. The higher the percentage is, the more hesitation the worker has, implying longer silence during the conversation.

In our study, *Topic cohesion* refers to whether the answers that workers give to predefined questions (described later in Section *Experimental Setup*) are topically coherent, and well linked. In some cases however, workers might directly ask questions to the conversational agent, referring to *4) Use of questions*, or express apologies to explain that they can not answer. Such questions or statements naturally deviate from the topic at hand. Therefore, we combine these two dimensions together as one factor in the coding scheme. The resulting set of dimensions used to systematically analyze conversation styles are summarized in Table 3.10, and they include: *1) Personal focus of topic*, *2) Enthusiasm*, *3) Pacing*, *4) Tolerance of silence*, and *5) Topic cohesion & Use of questions*.

Each dimension is quantified using a binary score (either -1 or 1). A final *score* is used to classify a conversation style as either *Involvement* or *Considerateness*. The score is calculated as a sum of scores corresponding to all the five dimensions. If final *score* is greater than 0, the conversational style of a worker is classified as *Involvement.* If the final *score* is less than 0, the conversational style of a worker is classified as *Considerateness.*

The coding scheme can be used to label ground truth data pertaining to conversational styles. To make the ground truth reliable, the coding process is carried out by multiple coders independently. Coders in the group independently score all the dimensions. The cases with disagreement are then resolved through manual discussion and disambiguation. The reliability of the coding process is measured by using Fleiss' Kappa [64].

### Conversational Style Prediction

We present two methods for conversational style prediction: one based on "rich conversation data" obtained from workers before microtask execution through a "pre-task conversation", and another based on "limited conversation data" obtained from task execution and a short "self-assessment". Figure 3.8 describes the organization of the envisioned conversational crowdsourcing task, where rich and limited conversation data are collected. Hereafter, we will refer to these methods as **Prediction with rich data** and **Prediction with limited data** respectively.

Table 3.11: Features used for conversational style prediction.

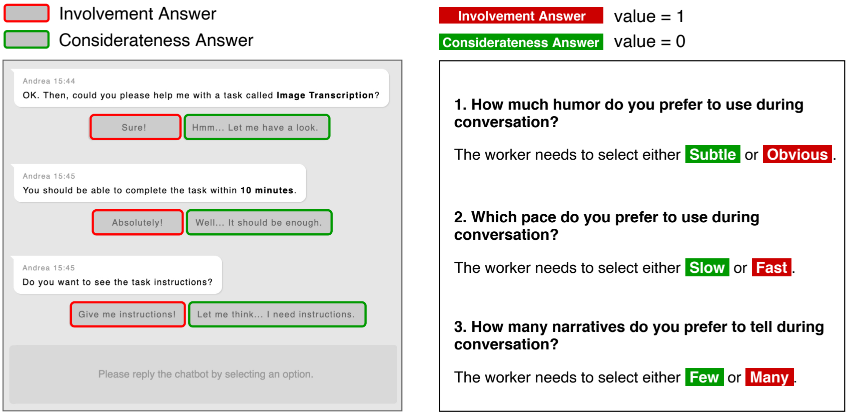
| ***Features*** | **Explanation** | **Related Dimensions** |
|---|---|---|
| *pp*† | percentage of personal pronouns. | personal focus of topic |
| *rep*† | mean repeat times of words. | topic cohesion |
| *wpu*† | mean number of words per utterance. | enthusiasm |
| *wps*† | mean number of words per sentence. | enthusiasm |
| *nque*† | number of question marks. | use of questions |
| *nexcl*† | number of exclamation marks. | enthusiasm |
| *pse*†* | percentage of self-editing. | tolerance of silence |
| *rtype*†* | rate of typing. | pacing |
| *pause*†* | mean pause length. | pacing, tolerance of silence |
| *tt*†* | mean length of turn-taking pause. | pacing, tolerance of silence |
| *nint*†* | number of interruptions. | pacing |
| *heu** | score from the heuristic assessment. | / |
| $sa_h$* | value of feature *humor* from self-assessment. | / |
| $sa_p$* | value of feature *pace* from self-assessment. | / |
| $sa_n$* | value of feature *narrative* from self-assessment. | / |

†= features used in **Prediction with rich data**.
* = features used in **Prediction with limited data**.

In case of **Prediction with rich data**, in the preliminary conversation the conversational agent initiates a brief discussion over an engaging topic, i.e. pre-task conversation. The resulting text is labeled as either *Involvement* or *Considerateness.* The coding scheme introduced in the previous section informed the design of several textual-features (as shown in Table 3.11), to be automatically extracted from the pre-task conversation. These features are used for training binary classifiers.

It is neither practical nor economically viable, to ask workers to engage in relatively long pre-task conversation with the agent before each task execution - a conversational style can be contextually dependent, so it cannot be considered as an immutable property of a worker. Therefore, we investigate the effectiveness of a **Prediction with limited data** method that does not require pre-task conversation, and needs a heuristic assessment and a short self-assessment instead.

*Heuristic assessment* takes place during the first three interactions of a task's execution, when the conversational agent is introducing the task title, time limit and task instructions respectively, as shown in Figure 3.7 (a). After each interaction, the agent provides the worker with two options to select. One option corresponds to Involvement style, while the other one corresponds to Considerateness style. The result *heu* derived from the heuristic assessment is also used as a feature for style prediction, which is calculated by $heu = h1 + h2 + h3$, representing the superimposition of answers from the first, second and third interactions respectively ($h_i$, $1 \leq i \leq 3$, is assigned to 1 if the Involvement answer is selected by the worker, otherwise it is assigned to -1).


🟥 Involvement Answer
🟩 Considerateness Answer
Sure!
Hmm... Let me have a look.
Absolutely!
Well... It should be enough.
Give me instructions!
Let me think... I need instructions.
Please reply the chatbot by selecting an option.
Involvement Answer value = 1
Considerateness Answer value = 0

Andrea 15:44
OK. Then, could you please help me with a task called **Image Transcription**?

Andrea 15:45
You should be able to complete the task within **10 minutes**.

Andrea 15:45
Do you want to see the task instructions?

(a) Heuristic-assessment

**1. How much humor do you prefer to use during conversation?**

The worker needs to select either Subtle or Obvious.

**2. Which pace do you prefer to use during conversation?**

The worker needs to select either Slow or Fast.

**3. How many narratives do you prefer to tell during conversation?**

The worker needs to select either Few or Many.

(b) Self-assessment

Figure 3.7: Explanations of (a) Heuristic-assessment and (b) Self-assessment of conversational style.

*Self-assessment* requires asking workers about their conversational styles. As shown in Figure 3.7 (b), we design a short self-assessment (that can be arranged either before or after crowdsourcing task execution as a part of pre- or post-task survey) indirectly asking workers about their preferences during the conversation, according to three continua used by Tannen in her example: *humor*, *pace* and *narrative*. Notice that these continua can be used in the scope of this self-assessment because they do not need to be inferred, but they can be directly reported by a user. The two optional answers of each question correspond to Involvement (value: 1) and Considerateness (value: 0) separately. Therefore three features pertaining to self-assessment ($sa_h$, $sa_p$ and $sa_n$ relate to humor, pace and narrative respectively) are considered in **Prediction with limited data**.

All the linguistic features for conversational style prediction (**Prediction with limited data** and **Prediction with rich data**) are listed in Table 3.11. Features *pp*, *rep*, *wpu*, *wps*, *nque*, and *nexcl* are only used in **Prediction with rich data**, because they are only available when the conversation allows for subjective elements (such as personal topics, narratives, opinions and emotions) to be expressed. Features *pse*, *rtype*, *pause*, *tt* and *nint* are used in both **Prediction with rich data** and **Prediction with limited data**, since they appear in any kinds of the conversation, including the ones for microtask execution. Features $heu$, $sa_h$, $sa_p$ and $sa_n$ come from the heuristic-assessment and the self-assessment, so they are only used in **Prediction with limited data**.

The linguistic features shown in Table 3.11 are automatically measured and used to train supervised binary classifiers for conversational style prediction. Each training example we need in this study is a pair consisting of an a vector of linguistic features (as input) and a binary output value (1 or 0, referring to Involvement and Considerateness respectively).

### Experimental Setup

To address **RQ3.3**, conversational styles are independently labeled by multiple coders according to the coding scheme to understand how workers' conversational styles distribute among crowd workers, and to investigate the feasibility of style prediction using rich conversation data and limited conversation data. To address **RQ3.4**, we analyze the relationship between workers' conversational styles and their performance, engagement, and cognitive task load. We used TickTalkTurk (Section 2.2) to deploy crowdsourcing tasks, which enables easy integration with existing platforms and access to the available crowd workers.

### Conversational Task Design

The conversational crowdsourcing task (HIT on MTurk) has three main phases, namely pre-task conversation, crowdsourcing microtasks, and post-task survey, as shown in Figure 3.8.

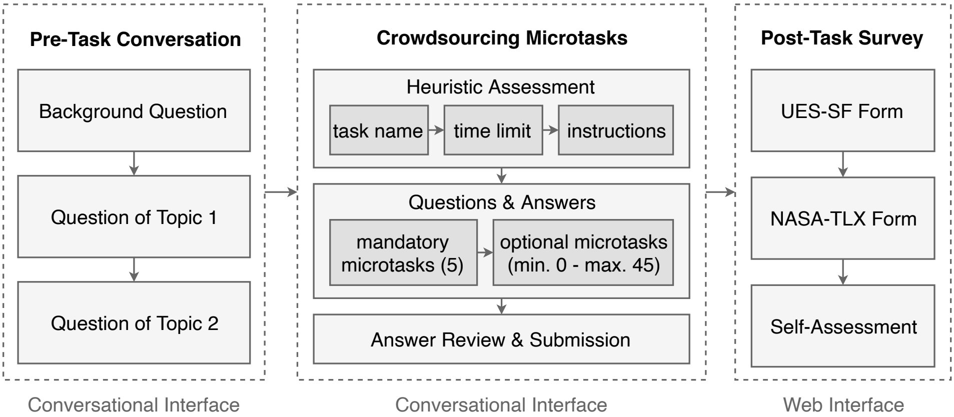


Figure 3.8: Organization of the conversational crowdsourcing task.

***Pre-task Conversation.*** As mentioned in Section 4, to acquire rich data for training classifiers, the conversational agent starts a pre-task conversation encouraging workers to share about their personal stories and opinions. The conversational agent asks three questions. The first question is about the demographic background of the worker. To facilitate the production of high-quality conversation data, the second and third questions are about two controversial topics (abortion and gun control respectively).

1 *Please tell me something about yourself (your age, your gender, your ethnicity and your education background). If you want, you can also tell me about your hobbies or interesting things you would like to share.*

2 *Do you think + [**Topic 1** in Table 3.12]?*

3 *Please explain what you think.*

4 *Do you think + [**Topic 2** in Table 3.12]?*

5 *Can you explain why?*

Table 3.12: Controversial topics used in pre-task conversation.

| ***Label*** | **Topic 1: Abortion** | **Topic 2: Gun Control** |
|---|---|---|
| *Pro* | abortion is getting rid of a fetus, not a human being | gun control guarantees safety of Americans |
| | reproductive choice empowers women | guns don't kill people, people kill people |
| | legalizing abortion helps to reduce sexual victimization | free access to guns is an important right |
| | social welfare systems cannot support unwanted kids well | guns make up only a small part of weapons that are used to commit crimes |
| | modern abortion is safe | people will always find a source for guns |
| *Con* | abortion is unsafe | guns are an important part of the US |
| | abortion is murder | we should control lunatics and criminals instead of guns |
| | abortion damages the well-being of the mother | banning guns will work not better than banning alcohol did |
| | women should not be able to use abortion as a form of contraception | armed guards or teachers will make schools safer |
| | women should accept the consequences of pregnant | gun control does not work |

The rationale behind this design is that controversial topics increase interest, which also increases the likelihood of conversation [32]. Although controversial topics have also been shown to increase discomfort [32], we prevented workers from diving into a deep discussion by asking only two questions per controversial topic. Thus, controversy in general is leveraged to better stimulate the desire of expressing opinions in our scenario. The content of these two questions are picked at random from those shown in Table 3.12, and are inspired from recent work by Hube et al. [102]. The corresponding labels of these two questions (also shown in the table) however, are not allowed to be the same simultaneously to avoid biases. Workers are required to provide at least three sentences (each sentence contains at least two words) for each question. If the worker does not meet this requirement, the conversational agent keeps prompting with – "Can you tell me more?", "Uh huh, and?", "Good, go ahead." or other similar phrases until it receives three sentences in total.

***Crowdsourcing Microtasks.*** The workflow of crowdsourcing microtasks on the conversational agent is also illustrated at the center of Figure 3.8.

First, the conversational agent provides workers with the basic information of the task, including task name, time limit, and task instructions. As we mentioned in the previous section, the interactions at the beginning are combined with a heuristic assessment (Figure 3.7).

After the worker selects their preferred options, the agent proceeds to the actual task execution part – questions & answers (i.e., answering microtasks, each microtask refers to a data row or a object). The worker has to complete 5 mandatory microtasks. After the mandatory part, the worker can choose either to stay or to leave. If a worker decides to stay (i.e., continuing task execution), at most 45 optional microtasks will be presented one after another until the worker asks the agent to stop the task execution. The number of answered optional microtasks is the quantitative measurement of worker retention in our study.

After the worker decides to stop the task execution, or to complete all the optional microtasks, the conversational agent sends an answer review to the worker to check if all the previous answers are correctly recorded by the agent. Finally, after the worker has reviewed and successfully submitted the answers, the agent will send a `Task Token` to the worker. Only with this `Task Token`, the worker can proceed to complete the post-task survey.

***Post-task Survey.*** User Engagement Scale Short Form [163, 164] (12 questions), NASA Task Load Index questionnaire[20] (6 questions) and the self-assessment of conversational style (3 questions) are used in the post-task survey to analyze worker engagement, cognitive task load, and conversational style.

First, workers have to complete the User Engagement Scale Short Form (UES-SF). The UES-SF consists of 12 questions in four factors (Focused Attention, Perceived Usability, Aesthetic Appeal, and Reward Factor). Workers are asked to answer each question by setting a slider on a 7-point sliding bar ranging from "*1: Strongly Disagree*" to "*7: Strongly Agree*".

Then, workers are asked to complete the NASA Task Load Index (NASA-TLX) questionnaire, which contains six items spanning six aspects (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration respectively). Workers need to indicate their perceived task loads on these six aspects by setting the slider to on a 20-point sliding bar ranging from "*1: Very Low*" to "*21: Very High*". The TLX scores will be scaled to 0-100 in the evaluation.

After UES-SF and NASA-TLX, workers are asked to complete the self-assessment of conversational style that has been described in Figure 3.7 (b).

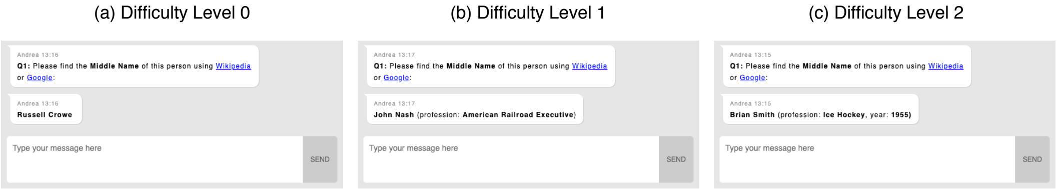
### Independent and Dependent Variables

***Independent Variables.*** Considering crowdsourcing tasks have different levels of difficulty and complexity [236], we design task into 3 difficulty levels, from easy to difficult, to observe how crowd workers with different conversational styles perform on different difficulty levels.

We consider two data types (text and image) of microtask, Information Finding and Image Transcription. We used these task types since they are representative of typical classes of microtasks [72], and they easily allow the classification of the task objective into different difficulty levels. This results in six experimental conditions (2 × 3), i.e. 2 types of task (Information Finding and Image Transcription, representing text-based and image-based tasks respectively) with three levels of difficulty (Level 0, Level 1 and Level 2, from easy to difficult).

[20] `https://humansystems.arc.nasa.gov/groups/TLX/`

*1) Information Finding Tasks* require workers to find the middle name of a famous person from either Wikipedia[21] or Google[22] [69]. As shown in Figure 3.9, this type of tasks has three different difficulty levels. In **Level 0**, the conversational agent shows first and last names of a unique, famous, and unambiguous person, whose middle names can be easily found from a search engine. In **Level 1**, the agent additionally shows the profession of the given person based on Level 0. We manually selected the names in this difficulty level to ensure there are at least two different famous persons with the same first and last names, so that the worker needs to distinguish the given person according the profession. In **Level 2**, the agent also shows the famous person's active year, based on Level 1 (showing first/last names and profession). In this difficulty level, there are multiple famous people with the same first/last name, and with the same profession. The worker is asked to find the correct one from those famous people by making use of all the given information.



(a) Difficulty Level 0

Andrea 13:16
**Q1:** Please find the **Middle Name** of this person using Wikipedia or Google:

Andrea 13:16
**Russell Crowe**

(b) Difficulty Level 1

Andrea 13:17
**Q1:** Please find the **Middle Name** of this person using Wikipedia or Google:

Andrea 13:17
**John Nash** (profession: **American Railroad Executive**)

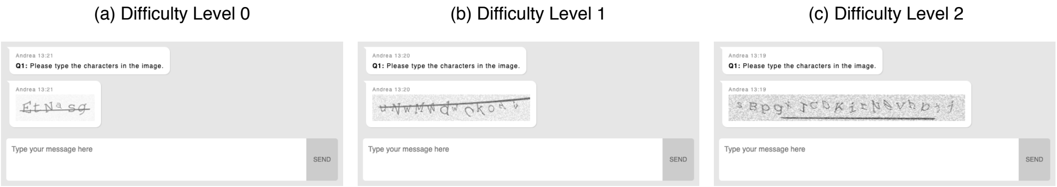
(c) Difficulty Level 2

Andrea 13:15
**Q1:** Please find the **Middle Name** of this person using Wikipedia or Google:

Andrea 13:15
**Brian Smith** (profession: **Ice Hockey**, year: **1955**)

Figure 3.9: Examples of the Information Finding task with three difficulty levels.

*2) Image Transcription Tasks* require workers to read the image (randomly generated by Claptcha[23]) and transcribe it into letters, as shown in Figure 3.10. This task type also has three different levels of the difficulty. In **Level 0**, the conversational agent shows an image containing a random combination of 6 letters (from the English alphabet, either in the upper or lower cases) with the highest resolution (approximately 38 px × 75 px per letter). The least noises (Gaussian white noises, $\sigma = 0.1$) are added into the image. In **Level 1**, the agent shows an image containing a random combination of 12 random letters with the medium resolution (35 px × 70 px per letter). The medium noises (Gaussian white noises, $\sigma = 0.3$) are added into the image. In **Level 2**, the agent shows an image containing a random combination of 18 random letters with the lowest resolution (approximately 33 px × 65 px px per letter). The largest noises (Gaussian white noises, $\sigma = 0.5$) are added into the image.



(a) Difficulty Level 0

Andrea 13:21
**Q1:** Please type the characters in the image.

Andrea 13:21

(b) Difficulty Level 1

Andrea 13:20
**Q1:** Please type the characters in the image.

Andrea 13:20

(c) Difficulty Level 2

Andrea 13:19
**Q1:** Please type the characters in the image.

Andrea 13:19

Figure 3.10: Examples of the Image Transcription task with different difficulty levels.

[21] http://en.wikipedia.org/

[22] http://www.google.com/

[23] https://github.com/kuszaj/claptcha

***Dependent Variables.*** We measure the performance of conversational style prediction, output quality, worker engagement and cognitive task load. We use these metrics to analyse what role the conversational style plays in microtask crowdsourcing.

*1) Performance of conversational style prediction* is measured by comparing prediction results of classifiers with the labeled ground truth produced based on the coding scheme. We measure the overall accuracy, noted as *accuracy*, which is the fraction of correctly predicted conversation style of workers among all the workers.

*2) Output quality* is measured using worker accuracy, which is calculated by comparing the answers provided by workers with the expected value:

$$\text{output quality} = \frac{\#\text{ correct answers provided by the worker}}{\#\text{ all the provided answers}} \tag{3.1}$$

In case of Information Finding tasks, the answer provided by workers should contain the expected value (case insensitive). In terms of Image Transcription tasks, the *string similarity* between the answer provided by workers and the expected value should be greater than 0.8. The string *similarity* is calculated as $similarity = 2 \times M/T$, where $T$ is the total number of characters in both answers and ground truth, and $M$ is the number of matched characters. Note that the *similarity* equals to 1 if the answer is identical to the ground truth (case insensitive), and equals to 0 if there is nothing in common (case insensitive).

*3) Worker engagement* is measured using two approaches: the first one is worker *retention*, quantified by the number of optional microtasks answered (from 0 to 45); and the second is the short-form of the user-engagement scale [164] – *UES-SF* scores in four different factors (Focused Attention, Perceived Usability, Aesthetic Appeal, and Reward Factor), ranging from 1 to 7. A relatively higher UES-SF score indicates that the worker is more engaged with regard to the corresponding factor [244].

*4) Cognitive task load* of workers is evaluated by using the unweighted NASA-TLX form, consisting of six questions. Workers are asked to give scores ranging from 0 to 100 to these questions. The final TLX score is the mean value of scores given to the six questions. Higher the TLX score is, the heavier task load the worker perceives.

### Experimental Environment

***Workers.*** There are 6 experimental conditions (2 task types × 3 difficulty levels), and each experimental condition has 50 crowdsourcing microtasks. As each microtask requires answers from at least 3 unique workers and each worker must submit at least 5 mandatory microtasks, we recruited $50 \times 3/5 = 30$ unique workers for each experimental condition from Amazon Mechanical Turk, resulting in $30 \times 6 = 180$ unique workers for the entire experiment. Only crowd workers whose HIT approval rates are greater than 95% could view and accept our crowdsourcing tasks [54].

***Rewards.*** To avoid monetary biases, we immediately pay 1 USD to each worker after the worker submits the task (HIT). Then all the workers equally receive 0.01 USD for each optional Image Transcription microtask, or receive 0.02 USD for each optional Information Finding microtask according to how many optional microtasks they answer after we approve their submissions. To ensure that we consistently pay an average hourly wage of 7.5 USD, we additionally bonus workers after calculating their execution time.

***Quality Control.*** To avoid biases caused by repeated task execution [69], we added extra Javascript code for recording their worker IDs on our server, to prevent workers from executing HITs multiple times. If a worker attempts to complete another HIT in this batch after having one already (meaning his/her worker ID has been recorded), all the instructions and questions on the web page are removed. Instead, a message, that kindly informs workers that they should "return" because of our experimental restrictions, is displayed on the task page.

## Evaluation

### Conversational Style Estimation

***Coding conversational styles.*** With this evaluation we address **RQ3.3**. The coding process was conducted by three coders who had deeply studied the theory of conversational style and understood the concept of linguistic devices. The inter-rater reliability was measured by Fleiss' Kappa. Three coders were in complete agreement for 124 out of 180 crowd workers. The 56 cases having disagreement were disambiguated manually by coders. In total, 86 workers exhibited *Involved* style, while 94 workers showed *Considerate* style. Therefore the kappa $\kappa$ value is 0.78.

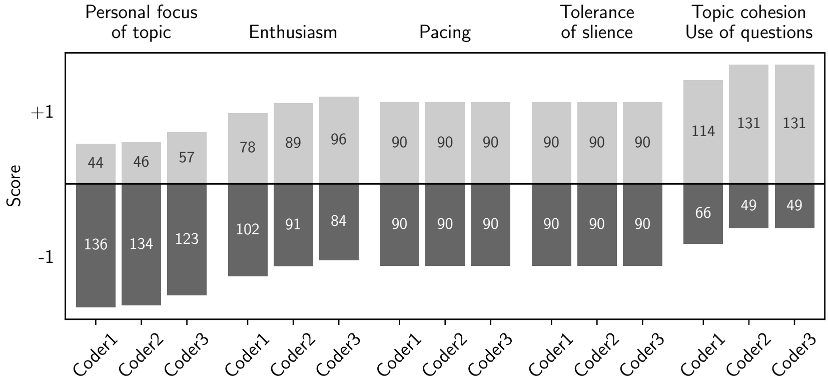


Figure 3.11: The score distribution of three coders across five dimensions. Numbers in this figure represent the number of judgments given by the corresponding coder.

The score distributions (of three coders, 180 judgments per dimension) of five dimensions are shown in Figure 3.11. *Pacing* and *Tolerance* were automatically calculated according to Table 3.10, therefore scores (-1 and +1) of these two dimensions are equally distributed. *Personal focus of topic*, *Enthusiasm* and *Topic cohesion & use of questions* were manually labeled by coders. As we can see from the Figure, Coder1 tended to give lower scores (more considerate) while Coder3 tended to give higher scores (more involved). However, scores given by different coders were distributed similarly in general (with only small fluctuations), suggesting that the coding scheme is sufficiently robust to account for the subjectivity of the coders.

***Prediction accuracy.*** The training dataset (features extracted by the conversational agent with ground truth data labeled by the coding scheme) was randomly divided into two part – 70% of them went to the training set, while the rest (30%) went to the testing set.

Because the focus of this study is the feasibility of the conversational style prediction, we did not use the state-of-the-art classifiers (e.g. deep neural network) and attempt to improve their performance. Instead, we only select most basic and naive classifiers (Logistic Regression, Naive Bayes, k-Nearest Neighbors, Decision Tree, Gradient Boosting), and compare the accuracy to understand the feasibility of predicting conversational styles.

The prediction performance is shown in Table 3.13. Gradient Boosting can obtain highest performance by using **Prediction with rich data**, whose overall accuracy value is nearly 80%. These results are encouraging, showing that it is indeed possible to achieve good prediction performance for conversational style using rich conversational data. In terms of **Prediction with limited data**, once again, Gradient Boosting can obtain highest performance, whose overall accuracy reaches 66.7%. These results are also encouraging, as they suggest that conversational styles can be predicted also through limited interactions with the workers. Since we didn't study deep into the parameter adjustment of machine learning models, we believe a well-tuned prediction method with limited data still has great potential to accurately estimate the conversational style.

Table 3.13: The accuracy (unit: percent %) of conversational style prediction by common classifiers, where the classifier with the highest accuracy is highlighted.

| ***Prediction methods*** | **Logistic Regression** | **Naive Bayes** | **k-NN** | **Decision Tree** | **Gradient Boosting** |
|---|---|---|---|---|---|
| *with rich data* | 75.93 | 75.93 | 74.07 | 77.78 | 79.63 |
| *with limited data* | 57.41 | 55.56 | 48.15 | 59.26 | 66.67 |

### The Performance of Workers with Two Conversational Styles

***Execution time.*** *Workers with an Involvement style generally featured less task execution time.* The average execution time of pre-task conversation, heuristic assessment & self-assessment, crowdsourcing microtasks, and USE-SF & NASA-TLX surveys are shown in Table 3.14. As workers with an Involvement style had a faster pace by definition, their task execution time were generally less than Considerate workers. We highlight that the average time spent on heuristic assessment and self-assessment is 73 seconds (around 0.15 USD per worker), while the average time spent on pre-task conversation is 412 seconds (around 0.86 USD per worker), where the latter is 5.6 times longer than the former, meaning Prediction with rich data is 5.6 times more expensive than Prediction with limited data.

***Output quality.*** *Workers with an Involvement style had significantly higher output quality at high difficulty level compared to Considerate workers.* We calculated the output quality (worker accuracy) across all types of tasks and difficulty levels, which are listed in Table 3.15. Obviously, the overall worker accuracy decreases as the task difficulty increases across two task types. We observed that when the overall worker accuracy was lower than 60% (IF Level 1, IF Level 2 and IT Level 2), the workers with *Involvement* style had higher average accuracy than the workers of *Considerateness* style. As the worker accuracy does not follow the normal

Table 3.14: The execution time ($\mu \pm \sigma$, unit: second) of pre-task conversation, heuristic assessment & self-assessment, crowdsourcing microtasks, and USE-SF & NASA-TLX surveys.

| ***Conversational styles*** | **Pre-task conversation** | **Heuristic- & self-assessment** | **Crowdsourcing microtasks** | **USE-SF & NASA-TLX** | **Total** |
|---|---|---|---|---|---|
| *Involvement* | $376 \pm 270$ | $69 \pm 55$ | $266 \pm 231$ | $129 \pm 179$ | $842 \pm 459$ |
| *Considerateness* | $444 \pm 247$ | $77 \pm 56$ | $318 \pm 297$ | $153 \pm 251$ | $994 \pm 551$ |
| *Overall* | $412 \pm 260$ | $73 \pm 56$ | $293 \pm 269$ | $142 \pm 220$ | $922 \pm 515$ |

distribution according to Shapiro-Wilk tests ($\alpha = 0.05$), with the Wilcoxon Rank-Sum test ($\alpha = 0.05$) we tested the significance of pairwise differences between two conversational styles. We found statistical significance for all the Level 2 Tasks (both Information Finding and Image Transcription, $p = 0.045$ and $p = 0.021$ respectively).

Table 3.15: The worker accuracy ($\mu \pm \sigma$: mean and standard deviation, unit: percentage) of all task types with difficulty levels.

| ***Task type*** | **Difficulty** | **Involvement** | **Considerateness** | **Overall** |
|---|---|---|---|---|
| *Information Finding* | Level 0 | $66.76 \pm 38.84$ | $73.77 \pm 33.17$ | $69.8 \pm 36.66$ |
| | Level 1 | $54.05 \pm 35.1$ | $50.42 \pm 35.22$ | $52.11 \pm 35.21$ |
| | **Level 2*** | $55.0 \pm 33.84$ | $25.95 \pm 28.84$ | $36.3 \pm 33.0$ |
| *Image Transcription* | Level 0 | $86.48 \pm 22.91$ | $90.56 \pm 8.31$ | $87.7 \pm 19.79$ |
| | Level 1 | $76.92 \pm 29.19$ | $79.41 \pm 26.0$ | $78.33 \pm 27.46$ |
| | **Level 2*** | $63.18 \pm 18.86$ | $35.68 \pm 32.74$ | $45.77 \pm 31.39$ |

* = statistically significant (Involvement vs Considerateness).

***Worker engagement: Worker retention.*** *We found no significant difference in worker retention.* We counted optional microtasks that workers answered, and plotted distributions of number of answered optional microtasks across different task types and difficulty levels using a violin plot (Figure 3.12). In this figure, the width of each violin at any point represents the number of workers that answered the corresponding number of optional microtasks. According to the shape of each "violin", the distributions do not meet any assumptions for parametric tests (also verified by Shapiro-Wilk tests), therefore we used the Wilcoxon Rank-Sum test (two-tailed, $\alpha = 0.05$) to test the significance of difference between two conversational styles. We found no significant difference in worker retention (answered optional microtasks) between the workers with two conversational styles.

***Worker engagement: UES-SF score.*** *Workers with an Involvement style reported significantly higher scores on UES-SF questionnaire in most cases of specific UES factors compared to Considerate workers.* UES-SF scores of four factors as well as overall scores are plotted in Figure 3.13. Since the distributions of UES-SF scores meet the assumption of normal distributions according to Shapiro-Wilk tests, to study user engagement of workers with different conversational styles, t-tests (two tailed, $\alpha = 0.05$) were performed to test the significance of differences between two conversational styles. Because multiple compar-

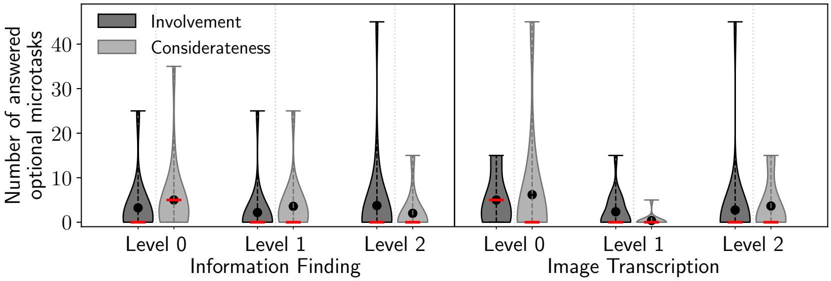


Figure 3.12: A violinplot representing the number of optional microtasks answered by workers across different task types and difficulty levels, where the red lines represent the median value and the black points represent the mean value.

isons (different UES factors) are conducted between two datasets (two conversational styles), Bonferroni correction is used to control Type-I error.

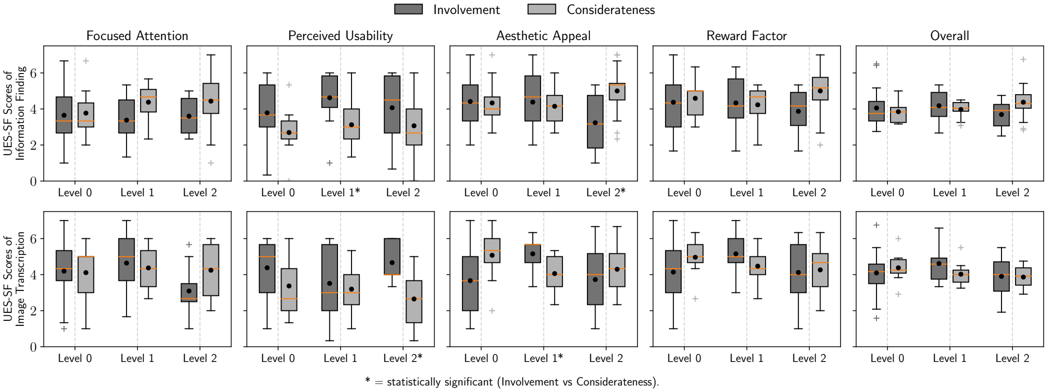
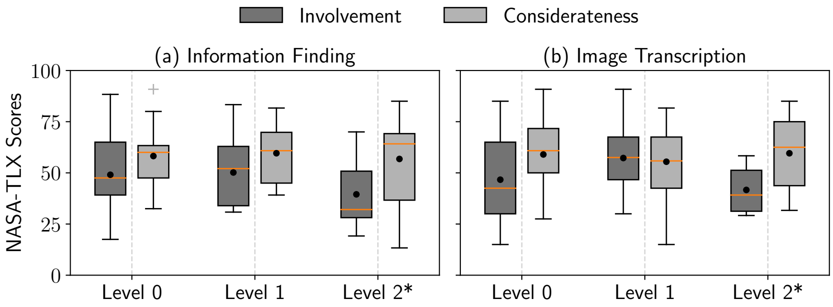


Figure 3.13: Boxplots of UES-SF scores by Task-Difficulty and Conversational Style, where the red lines represent the median value and the black points represent the mean value.

In terms of Information Finding tasks, whose UES-SF scores are displayed in the first row of Figure 3.13, workers of Involvement style reported significantly higher scores with respect to Perceived Usability factor, when difficulty level is 1 (IF Level 1 on Perceived Usability, $p = 0.0033$). Workers of Considerateness style reported higher Aesthetic Appeal score at level 2 (IF Level 2 on Aesthetic Appeal, $p = 0.0026$). As for Image Transcription tasks (UES-SF scores are displayed in the second row of Figure 3.13), workers of Involvement style gave higher scores on Perceived Usability and Aesthetic Appeal when difficulty level is 2 and 1 respectively (IT Level 2 on Perceived Usability and IT Level 1 on Aesthetic Appeal, $p = 0.0003$ and $p = 0.0034$ respectively). We found no significant difference between two styles in terms of Overall UES-SF score.

To conclude, workers with an Involvement style tended to report higher scores on UES-SF questionnaire, while workers with both styles did not show significant differences with respect to worker retention in both tasks.

***Cognitive task load.*** *Workers with an Involvement style reported significantly less cognitive task load at high difficulty level compared to Considerate workers.* Results of unweighted NASA-TLX scores are displayed as box plots in Figure 3.14. As we can see from the figure, workers of Involvement style reported lower mean values than workers of Considerateness style in all the experimental conditions except only one case (IT Level 1). Similarly, those Involvement workers also reported lower median values across all the experimental conditions except the same case (IT Level 1).



* = statistically significant (Involvement vs Considerateness).

Figure 3.14: Boxplots of NASA-TLX scores by Task-Difficulty and Conversational Style, where the red lines represent the median value and the black points represent the mean value.

Since the distributions of TLX scores meet the assumption of normal distributions according to Shapiro-Wilk tests, we conducted t-tests (two tailed, $\alpha = 0.05$) to find significant pairwise differences between two conversational styles across different task types and difficulty levels. Results show that workers of Involvement style reported less cognitive task load than workers of Considerateness style, after they completed all the tasks of Level 2 with significant differences (both Information Finding and Image Transcription, $p = 0.034$ and $p = 0.006$ respectively). These results collectively suggested that workers of Involvement style perceived less task load from task execution than workers of Considerateness style in general, especially when completing difficult tasks.

## Discussion

### Reflection on Conversational Style Estimation

We proposed a coding scheme for conversational style estimation that systematically classifies the text-based conversational style into two categories – *Involvement* and *Considerateness*. To the best of our knowledge, this is the first work that proposes a systematic method to estimate the conversational style of online workers. We also studied the feasibility of automatically predicting workers' conversational styles by common classifiers using rich conversational data (more accurate) and limited conversational data (less expensive) respectively. Results showed that indeed conversational styles could be predicted, using the proposed task design, thus showing that conversational style estimation could serve as a practical tool for microtask crowdsourcing.

### Reflection on the Performance of Workers

We explored the behavior of online workers with two conversational styles during conversational microtasking, and observed strong evidence that conversational style could bear relationship with quality of outcome for difficult tasks. We found that workers with an Involvement style performed better in terms of quality-related outcomes in tasks with higher difficulty levels. These results suggested that conversational style estimation could be a useful tool for output quality prediction. Analysis of cognitive task load revealed that workers with an Involvement style perceived less task load with higher difficulty levels. Our experimental findings suggested that the conversational style estimation could be used for worker performance prediction to better enable adaptive crowdsourcing strategies.

## Design Implications

The results showed that the conversational style had a significant impact on output quality, worker engagement, and cognitive task load. Workers with an Involvement style could produce higher work accuracy, perceive higher user engagement and feel less cognitive task load when they were completing difficult tasks. This gives us a strong indication that conversational style estimation could be an effective tool for predicting worker performance and assisting crowdsourcing task assignment. Highly involved workers could be selected and assigned to tasks of higher complexity, to produce high-quality work outcomes. The proposed method can be applied in microtask crowdsourcing marketplaces to improve work outcomes and foster a better worker-requester relationship (by improving worker accuracy, increase worker engagement, and reducing cognitive task load).

We found that a long conversation (pre-task conversation) could provide enough data for precisely (80%) predicting the conversational style, however, it took 5.6 times longer than a heuristic assessment with a self-assessment (67% and only took around 1 minute). The precision of prediction with limited data was lower but acceptable, and we are confident that it could be improved using state-of-art classifiers. We suggest that in the future design of the conversational microtask crowdsourcing, a heuristic assessment with a self-assessment could be considered as an extra test to effectively predict worker performance (which also help in dynamically adapting task scheduling and assignment strategies), while it only costs 0.1 to 0.2 USD additionally on each worker.

In this study, we propose a conversational style estimation method for conversational microtask crowdsourcing. Since previous studies have shown the feasibility of deploying microtasks on Facebook (Messenger) [85, 22, 218], Twitter [193], etc., the proposed method can be applied in chatbot systems on common social network channels for analyzing and understanding the personality, mood, subjectivity, and bias of (particularly anonymous) users.

The estimation and prediction of the conversational style of a worker has clear implications for privacy and fairness. While performing a dialogue, workers could disclose personal information that should be treated according to best practices and regulation for personal data management. In this respect, the *Prediction with limited data* allows for a data collection method that is minimally invasive and that could be easily repeated across HITs, thus eliminating the need for storage and management of worker data while achieving good prediction performance. We also stress the potential fairness implications that conversational style profiling can have on task distributions and therefore, on the earning power of workers. We believe that user modelling approaches like the one we propose should be used only in the context of an explicit, optional, and informed approval from workers, and such that a lack of approval does not lead to overall compensation penalties.

### Limitations and Future Work

In terms of style estimation and prediction, we only focused on the text-based conversation. Text-based conversation ignores several paralinguistic features (pitch, voice) and nonlinguistic features (smile, laughter, gestures). Moreover, some features relying on the analysis of voice such as pacing were measured in a different way. In general, there are various means to interact with conversational agents (e.g., voice-based agent, video-based agent). Conversational agents and corresponding style estimation methods based on voice or video could be an interesting direction to explore. Second, there is still room for improvement for style prediction. In this study, we used the most basic and naive binary classifiers for style prediction, and did not tune the parameters of those classifiers to pursue higher accuracy. Thus, in the imminent future, state-of-the-art machine learning methods can be applied for conversation style prediction. The process of labeling ground truth data by using the coding scheme can also be crowdsourced, to label larger training datasets with the cooperative work of crowd workers.

In terms of the study of the impact of conversational styles, the task types studied in the experiment are limited. We only focused on one input type of microtask – that is free text. In the case of this input type, workers were asked to provide their responses using free text (string). However, many other input types such as multiple choices, sliding bars, and bounding boxes are also used in microtasks of current online crowdsourcing marketplaces. Studying the performance of crowd workers with different conversational styles on other types of tasks is an important next step to our work. Previous work shows that an aligned style of the conversational agent can improve worker performance [195, 215]. Future experiments should consider assigning different conversational styles to the conversational agent, and investigate whether an aligned style can help in the improvement of worker engagement and reduction of cognitive task load.

Moreover, to estimate conversational styles, the crowd workers in our experiments spent a long time on pre-task conversations (around 7 minutes), which might have a negative impact on work outcomes. A future study can explore style prediction and alignment on the experimental conditions without additional conversation to further evaluate the breadth of practical value in conversational style prediction. Furthermore, future work can focus on the usage of a variety of different types of conversational agents, the collection of large amount of conversational data for constructing a training dataset and deep-learning-based classifiers for conversational style estimation.

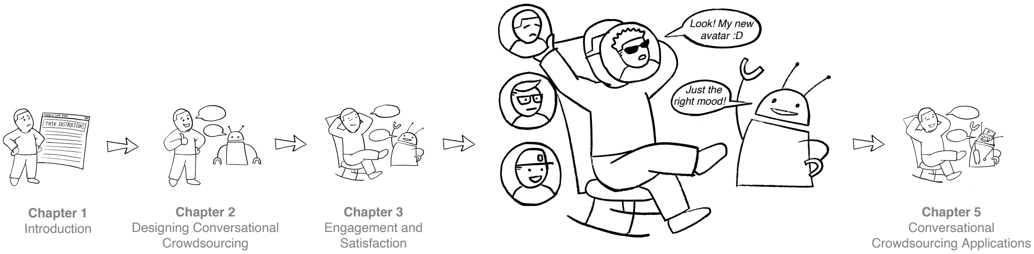
## 3.3 Chapter Summary

We studied the impacts of conversational crowdsourcing and the use of conversational styles on worker satisfaction and engagement in this chapter. In the first study (Section 3.1), we conducted online crowdsourcing experiments to study whether the worker engagement can be affected by the conversational interface. We measured workers' user engagement and cognitive load while completing tasks using conversational interfaces with different conversational styles. In the second study (Section 3.2), we proposed a coding scheme for style estimation based on the five dimensions of examined linguistic devices and style prediction methods. We performed a crowdsourcing experiment to analyze the behaviour of crowd workers with different conversational styles.

We show that the use of conversational crowdsourcing can effectively improve the perceived worker engagement and worker retention. We also highlight that the conversational style has a significant impact on output quality, worker engagement, and cognitive task load. The coding scheme proposed in this chapter can reliably estimate conversational styles of crowd workers, which could be used as an effective tool to evaluate and predict worker performance, and further assist in designing dynamic and personalized crowdsourcing strategies. The two studies carried out in this chapter provide important insights in terms of improving worker satisfaction and engagement in crowd work.

# Chapter 4

# The Roles of Mood and Self-Identification



In the previous chapters, our studies have revealed the potential of conversational crowdsourcing in improving worker engagement and satisfaction. Compared to traditional web crowdsourcing, there is a stronger link between conversational crowdsourcing and worker emotions, due to the human-like interaction of conversational interfaces. However, how the emotions of crowd workers can affect conversational crowdsourcing remains unanswered. Lately, related literature revealed that worker moods have been shown to have significant effects on quality-related outcomes in the context of crowdsourcing. Meanwhile, self-identification is found to be strongly associated with emotions in the realm of games, which has received attention in recent HCI literature. Particularly, recent work has shown that self-identification with player avatars is effective in fostering interest, enjoyment, and other emotional aspects pertaining to intrinsic motivation. However, little is known about the role of worker moods and self-identification in shaping work in conversational crowdsourcing. Therefore, in this chapter, we carried out two studies to understand the roles that worker moods and self-identification could play in conversational crowdsourcing.

In the first study (Section 4.1), we conducted a crowdsourcing study addressing 600 unique online workers, to investigate the role that worker moods could play in conversational crowdsourcing. We also explored whether suitable conversational styles of the agent could affect the performance of workers in different moods. Our results showed that workers in a pleasant mood tended to produce significantly higher quality results, exhibit greater engagement and report a lower cognitive load, and a suitable conversational style could have a significant impact on workers in different moods.

In the second study (Section 4.2), we carried out a between-subject study involving 360 crowd workers. We investigated how worker avatars influence quality related outcomes of workers and their perceived experience, in conventional web and novel conversational interfaces. We equipped workers with the functionality of customizing their avatars, and selecting characterizations for their avatars, to understand whether identifying with an avatar could increase the motivation of workers. We found that using worker avatars with conversational interfaces could effectively reduce cognitive workload and increase worker retention. Our results indicated the occurrence of similarity and wishful avatar identification in conversational crowdsourcing.

The two studies carried out in this chapter investigated how workers' subjective perceptions could affect conversational crowdsourcing, and showed the important roles that worker mood and self-identification could play. Our findings advance the current understanding of conversational crowdsourcing, and have important implications in improving worker subjective experience and on the design of future conversational crowdsourcing systems.

The content of this chapter is based on the following papers:

Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon. Just the Right Mood for HIT! Analyzing the Role of Worker Moods in Conversational Microtask Crowdsourcing. International Conference on Web Engineering, pp. 381-396, 2020. (Section 4.1 is based on this paper)

Sihang Qiu, Alessandro Bozzon, Max V. Birk, Ujwal Gadiraju. Using Worker Avatars to Improve Microtask Crowdsourcing. Proceedings of the ACM on Human-Computer Interaction (CSCW), pp. 1-28, 2021. (Section 4.2 is based on this paper)

## 4.1 The Role of Worker Moods

In the previous chapters, we have argued that conversational crowdsourcing had advantages over traditional graphical user interfaces since it can be used to improve worker engagement and satisfaction. However, to advance the understanding of conversational crowdsourcing, it is worth exploring how workers' subjective perceptions can affect conversational crowdsourcing. Worker *moods* are known to influence the quality of work in the general workplace [217], including online microtasking platform where microtasks are executed using traditional web interfaces [235, 244]. For example, workers in a happy mood were found to exhibit a better performance than those who were less happy [233, 241]. Others have shown that worker moods can also impact task execution time [151]. Recent work in the context of online crowdsourcing has revealed the relationship between worker moods and crowdsourcing task performance [244], where moods were measured using the Pick-A-Mood instrument [48] and statistical tests indicated that worker moods had significant effects on their engagement. Based on these findings, others analyzed the impact of worker moods in struggling web search tasks [67]. There is a limited understanding however, of how moods of workers interact with conversational interfaces in shaping the quality of their work. Furthermore, an opportunity to improve conversational conversational further, lies in analyzing the potential impact of conversational styles [211] of agents on quality related outcomes of workers in different moods. In Chapter 3, we have investigated whether adapting and personalizing the conversational style of an agent to that of a worker can improve the quality of work. To this end, we explore whether a conversational agent with different conversational styles can enable workers in different moods to produce better task performance or to have better microtasking experience. We aim to fill this knowledge gap by addressing the following research questions:

**RQ4.1:** How do worker moods affect their performance, engagement and cognitive load in conversational crowdsourcing?

**RQ4.2:** How does the conversational style of a conversational agent affect the performance of workers in different moods?

We used the conversational interface with different conversational styles designed and implemented in Section 3.1 to support workers in the execution of Human Intelligence Tasks, i.e. HITs. We carried out a crowdsourcing study with 600 unique workers, across four types of tasks and three different interfaces (3 × 4 = 12 experimental conditions in total). To answer **RQ4.1**, we evaluated the performance of workers, their engagement (using the User Engagement Scale-*UES*) and cognitive load (NASA-TLX) across different tasks. Results revealed that workers in a pleasant mood tended to produce significantly higher quality results (over 20% improvement), exhibited greater engagement (over 18% improvement) and reported a lower cognitive load (a decrease by nearly 13%). To address **RQ4.2**, we considered three different interfaces (traditional web interface, and conversational interfaces with two conversational styles). Results demonstrated that a suitable conversational style could have a significant impact on workers in terms of their engagement and cognitive task load.

# Method

## Workflow and Task Design

The entire task execution process across different conditions consists of four main stages: self-reported mood (Pick-A-Mood), a short demographic survey, the crowdsourcing HITs, and a post-task survey, as illustrated in Figure 4.1.

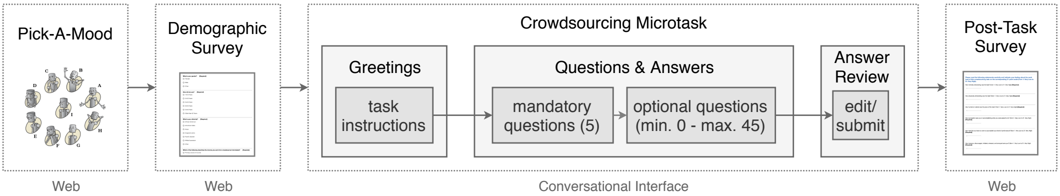


Figure 4.1: Crowdsourcing microtask workflow in the conversational interface conditions.

*1) Pick-A-Mood.* Workers are first asked to self-report their moods using the Pick-A-Mood instrument shown in Figure 4.2. Nine moods are presented, and can be grouped into three categories, which are **pleasant** moods (A: *cheerful*, B: *excited*, H: *relaxed* and G: *calm*), **unpleasant-** moods (C: *tense*, D: *irritated*, E: *sad* and F: *bored*) and a **neutral** mood (I).

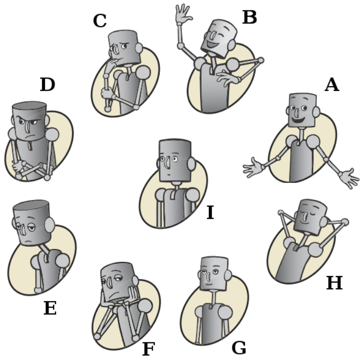


Figure 4.2: Pick-A-Mood – a self-reported scale to measure the mood of crowd workers.

*2) Demographic Survey.* Next, workers are asked to respond to simple background questions pertaining to their gender, age, ethnicity, educational background, and sources of income.

*3) Crowdsourcing HIT Design.* The actual crowdsourcing HITs are executed on either the conversational interface or the traditional web interface as per the experimental condition. The microtasks batch has 5 mandatory HITs and 45 optional HITs. Workers must complete the 5 mandatory HITs to proceed to the next stage. On completing the mandatory HITs in the conversational interface condition(s), the agent asks the workers if they want to continue on and complete more HITs. In case of the traditional web interface condition(s), workers can click a button named '`I want to answer more questions`' to complete more optional HITs.

*4) Post-task Survey.* The last stage of the workflow presents workers with a survey, to gather the worker's perception about the HITs completed. Workers are first asked to complete the User Engagement Scale Short Form [163, 164] (UES-SF). Within this, 12 questions need to be answered by adjusting the slider bar ranging from "*1: Strongly Disagree*" to "*7: Strongly Agree*". O'Brien designed the UES for systematically measuring user engagement through self-assessment [163], and later developed the short form of UES (UES-SF) to be suitable for time-sensitive contexts [164]. Next, workers are asked to complete the NASA Task Load Index (NASA-TLX) questionnaire[24], which includes six questions corresponding to different kinds of cognitive task load (ranging from "*0: Very Low*" to "*100: Very High*").

### Conversational Interface

To support the execution of HITs on a conversational interface, we incorporate the following aspects.

*1) Greetings.* Drawing from the essential structure of conversation, the conversational interaction begins with greetings. The goal here is to let workers familiarize themselves with the conversational interface. Next, the conversational interface then helps workers understand how to execute HITs by introducing the task instructions using dialogues.

*2) Questions & Answers.* The conversational interface asks questions to workers, and workers can answer these questions by either typing answers or using provided UI (user interface) elements.

*3) Answer Review.* On the traditional web interface, a worker can easily go back to a question and edit its answer. To realize this affordance in the conversational interface, workers are provided with the opportunity to review and edit their answers if needed, before submitting the HITs.

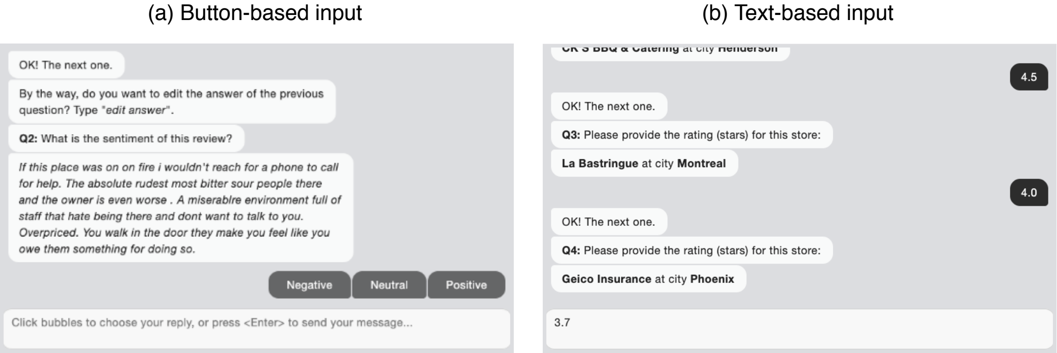
The user interfaces of most common crowdsourcing platforms mainly support HTML/CSS and Javascript. To make sure the conversational interface can be directly embedded into such platforms, we used the conversational interface developed in Section 3.1, based on a HTML/Javascript chatbot project `chat-bubble`[25]. This allows us to avoid redirecting workers to an external chatting or messaging application. The conversational interface supports two modes of input – free text and multiple choices, since these two types of input can enable workers to effectively provide judgments for most popular crowdsourcing task types [66]. As shown in Figure 4.3, `bubble-like buttons` and `textarea` (at the bottom of UI) are used for supporting the input modes of multiple choice selection and free text entry respectively.

### Conversational Style

We also investigate whether a suitable conversational style of the conversational agent can affect the performance of workers in different moods. As we studied in Chapter 3, according to Deborah Tannen's seminal theory, conversational styles can be classified into two broad categories, namely *High-Involvement* and *High-Considerateness* [210, 211]. A conversational style is actually the superimposition of multiple linguistic features and devices. To this end, we used the design criteria from Section 3.1 (Table 3.1) to create conversation agents emulating High-Involvement and High-Considerateness conversational styles. Table 4.1 shows

[24] `https://humansystems.arc.nasa.gov/groups/TLX/`

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



(a) Button-based input

OK! The next one.

By the way, do you want to edit the answer of the previous question? Type "*edit answer*".

**Q2:** What is the sentiment of this review?

*If this place was on on fire i wouldn't reach for a phone to call for help. The absolute rudest most bitter sour people there and the owner is even worse . A miserablre environment full of staff that hate being there and dont want to talk to you. Overpriced. You walk in the door they make you feel like you owe them something for doing so.*

Click bubbles to choose your reply, or press <Enter> to send your message...

(b) Text-based input

**CK'S BBQ & Catering** at city **Henderson**

4.5

OK! The next one.

**Q3:** Please provide the rating (stars) for this store:

**La Bastringue** at city **Montreal**

4.0

OK! The next one.

**Q4:** Please provide the rating (stars) for this store:

**Geico Insurance** at city **Phoenix**

3.7

Figure 4.3: Conversational interfaces for execution of HITs provide two input means: (a) buttons and (b) free text.

examples of how the conversational agent opens a conversation while emulating the two different conversational styles.

Table 4.1: Examples of greetings with High-Involvement and High-Considerateness styles.

| **High Involvement** | **High Considerateness** |
|---|---|
| — *Hey! Can you help me with a task called Information Finding?* | — *Thank you in advance for helping me with a task called Information Finding.* |
| — *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.* |
| — *Here is the task instructions. Take a look!* | — *I kindly ask you to have a look at the task instructions.* |

## Experiments and Setup

### Experimental Design

In our experiments, we consider two data types (image and text) and two input types (free text and multiple choices), resulting in 4 HIT types (2 data types × 2 input types) - Information Finding (text data + free text input), Sentiment Analysis (text data + multiple choices), CAPTCHA Recognition (image data + free text input) and Image Classification (image data + multiple choices). The experiment is approved by the ethics committee of TU Delft.

In **Information Finding (IF)** tasks, workers are asked to find and provide the rating (stars) of a given store from Google Maps. In **Sentiment Analysis (SA)** tasks, workers are asked to read given reviews of stores and determine the overall sentiment of the review. In **CAPTCHA Recognition (CR)** tasks, workers are asked to observe the image and determine which letters the image contains, and then provide the letters in the same order as they appear in the image. In **Image Classification (IC)** tasks, workers are asked to determine which animal the image contains.

We consider three distinct interfaces: **1) Traditional web interface (Web)** where all the HITs are displayed and answered using traditional HTML elements; **2) Conversational interface with High-Involvement style (Con+I)**, where the HITs are presented through an agent with a High-Involvement style; **3) Conversational interface with High-Considerateness style (Con+C)**, which is similar to Con+I, except that the agent converses with workers using a High-Considerateness style.

Thus, the four task types and three interfaces result in a cross-section of 12 experimental conditions. These 12 experimental conditions were published on Amazon Mechanical Turk (MTurk) as HIT batches in our experiments.

### Evaluation Metrics

The evaluation metrics in our experiments are *output quality*, *worker engagement*, and *cognitive task load.*

*Output quality* is measured using the accuracy of workers. A worker's accuracy is calculated as the fraction of correct responses over the total number of responses provided by a worker. Here, we consider a HIT to be accurately completed if and only if the response is identical to the ground truth (case insensitive).

*Worker engagement* is measured using: 1) worker retention, quantified by the number of optional HITs completed (ranging from 0 to 45); and 2) the UES-SF scores ranging from 1 to 7. A higher UES-SF score indicates that the worker is relatively more engaged.

*Cognitive task load* is evaluated by unweighted NASA-TLX form, consisting of six questions. Workers are asked to give scores ranging from 0 to 100 to these questions. The final TLX score is the mean value of scores given to the six questions. The higher the TLX score is, the greater is the task load perceived by a worker.

### Workers and Rewards

In our setup, each experimental condition consists of 50 HITs and we recruited 50 unique workers to participate and complete the workflow in each case. As a result, we acquired judgments from $12 \times 50 = 600$ unique workers.

After a worker provided a valid `task token` and successfully submitted the HITs on MTurk, the worker was immediately paid 0.5 USD, a fixed payment for successful submission. To reach an average hourly wage of 7.5 USD, we provided bonuses to workers according to the number of optional HITs that they completed. Workers working on image-based tasks (CAPTCHA Recognition and Image Classification) received 0.01 USD for each optional HIT, while workers working on text-based tasks (Information Finding and Sentiment Analysis) received 0.02 USD for each optional HIT.

### Quality Control

Although MTurk allows task requesters to set a qualification type to prevent workers from executing tasks in multiple HIT batches, workers are still able to execute multiple HITs from a single batch. To ensure each worker at most submits once, we recorded unique worker IDs on our server using Javascript, to prevent repeated participation. To ensure reliability of results, validity of responses, and control for potential malicious activity [54, 74], we restricted participation by using an MTurk qualification attribute, only allowing crowd workers whose HIT approval rates were greater than 95% to access our tasks.

## Results

### Worker Demographics

Of the unique 600 workers, 36.6% were female and 63.4% were male. The majority of workers were found to be Asian (46.37%), while 39.12% of workers were Caucasian. Most workers (89.2%) were under 45 years old, and education levels of most workers (74.5%) were higher than (or equal to) Bachelor's degree. In terms of source of income, 38.0% of the workers claimed MTurk was their primary source of income, while 55.4% of the workers worked on MTurk part-time and considered it as their secondary source of income. We publicly released all data (HITs deployed and responses from workers across the different experimental conditions) to facilitate further research for the benefit of the community[26].

### Distribution of Worker Moods

According to the results from the Pick-A-Mood instrument, 74.45% of workers reported to be in a pleasant mood, and 22.67% of workers reported unpleasant moods. Only 2.88% of workers reported to be in a neutral mood. As shown in section 4.1, most workers reported to be in a cheerful mood. Consistent with prior findings in microtasking marketplaces [67, 235, 244], we found that a majority of workers were in pleasant moods.

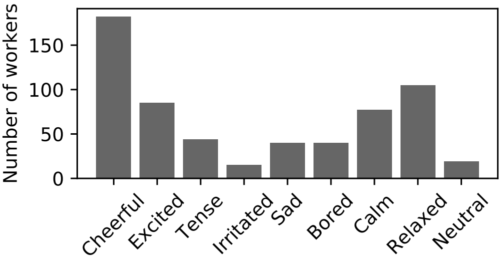


Figure 4.4: Overall distribution of worker moods.

Figure 4.1 shows the distribution of worker moods across all experimental conditions, where IF, SA, CR and IC represent Information Finding, Sentiment Analysis, CAPTCHA Recognition, and Image Classification respectively. Web, Con+I and Con+C refer to the web interface, conversational interface with involvement-style and conversational interface with considerateness-style in each case. The mood distribution of workers within each experimental condition is similar to the overall mood distribution. Moreover, there were no workers who reported a neutral mood in web interface conditions of Information Finding and Sentiment Analysis tasks, and the conversational interface with High-Considerateness style of Information Finding (IF Web, IF Con+C and SA Web). Since there were only a few workers with a neutral mood who executed HITs across different experimental conditions, we excluded the workers in a neutral mood in our analysis presented further.

[26] Companion page: `https://sites.google.com/view/icwe2020mood`

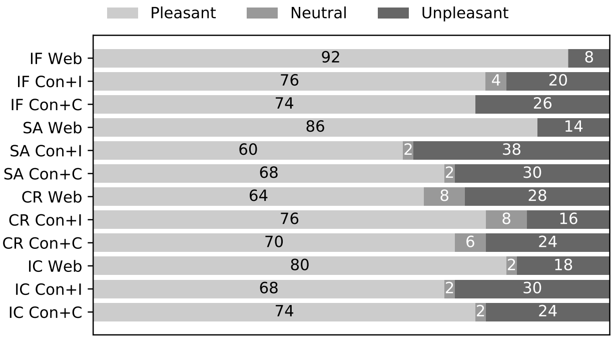


Figure 4.5: Percentages of workers in pleasant, neutral and unpleasant moods across different experimental conditions.

### Worker Performance

We analyzed the performance of workers across different experimental conditions. Figure 4.6 shows the output accuracy of workers. Due to the relative ease of tasks, in case of image-based HITs (CAPTCHA Recognition and Image Classification), the output accuracy of workers is generally higher and more stable across different interfaces and worker moods, compared to that in text-based HITs (Information Finding and Sentiment Analysis).

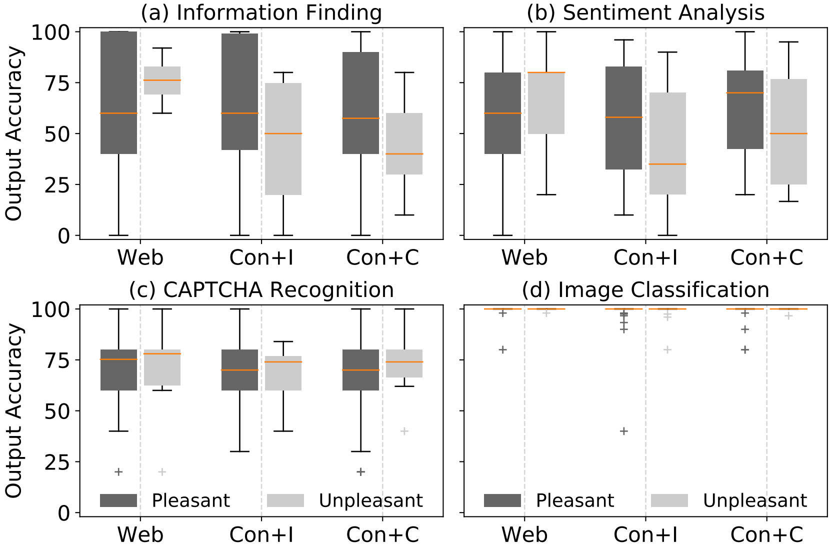


Figure 4.6: Boxplots showing the output accuracy (unit: %) of workers in different moods, across different experimental conditions. Red lines in boxplots indicate the median value.

To assess whether moods can affect worker performances in different interfaces, we conducted t-tests (two-tailed, $\alpha = 0.05$) to test the significance of pairwise differences between different interfaces within one conversational style. Results show that the performance of workers in unpleasant moods, using the conversational interface with High-Considerateness style (Con+C, $\mu = 43.1$, $\sigma = 23.0$) is significantly lower than those using the web interface (Web, $\mu = 76.1$, $\sigma = 11.6$) in Information Finding task (unpleasant, IF Con+C vs. IF Web, $p = 0.02$). In general, we found that the output quality corresponding to workers in unpleasant moods using conversational interfaces (both Con+I and Con+C) is generally lower than those using the traditional web interface on text-based tasks. This can intuitively be explained by the potential aversion of workers to engage with a conversation when in an unpleasant mood [118].

To investigate how workers with different moods perform under the same condition, we tested the statistical differences between the performance of workers across the two conversational styles using t-tests (two-tailed, $\alpha = 0.05$). Workers in pleasant moods performed significantly better than those in unpleasant moods, while using conversational interfaces with High-Involvement (pleasant $\mu \pm \sigma = 68.2 \pm 28.0$ vs. unpleasant $\mu \pm \sigma = 46.3 \pm 28.6$) and High-Considerateness styles (pleasant $\mu \pm \sigma = 63.3 \pm 29.8$ vs. unpleasant $\mu \pm \sigma = 43.1 \pm 23.0$) for executing Information Finding HITs (pleasant vs. unpleasant on IF Con+I and IF Con+C, $p = 0.031$ and $p = 0.033$ respectively). In general, our results suggest that workers in pleasant moods exhibited a higher quality while using conversational interfaces, in comparison to workers in unpleasant moods.

### Worker Engagement

***Worker Retention.*** Figure 4.7 shows the number of optional questions that workers answered across different task types, interfaces and moods. Since the number of optional HITs completed does not follow a normal distribution, we conducted Wilcoxon rank-sum tests (two-tailed, $\alpha = 0.05$) to test for statistical significace.

By comparing worker retention of different moods within each experimental condition, we found that the retention of workers in pleasant moods ($\mu = 7.2$, $\sigma = 10.7$) is significantly lower than that of workers in unpleasant moods ($\mu = 10.8$, $\sigma = 8.1$) using conversational interfaces with the Considerateness style for executing the Sentiment Analysis HITs (pleasant vs. unpleasant on SA Con+C, $p = 0.027$). This suggests that conversation interfaces with a particular conversational style can have the potential to improve worker retention based on the task type.

We found that workers in pleasant moods using conversational interfaces (both High Involvement and High Considerateness, Con+I and Con+C) answered significantly more optional HITs than workers in pleasant moods using traditional web interfaces across all four types of tasks (pleasant, all task types, $p < 0.05$). Workers in unpleasant moods also answered more optional HITs using conversational interfaces (both Con+I and Con+C) than those using web interfaces in Sentiment Analysis and CAPTCHA recognition with significant differences (unpleasant, SA and CR, $p < 0.05$).

***User Engagement Scale (UES-SF).*** We aggregated and analyzed the responses of workers in the post-task survey. Figure 4.8 depicts the UES-SF scores of workers across all types of tasks, interfaces and two different moods (pleasant vs. unpleasant). To understand the effect of worker moods on user engagement, t-tests (two tailed, $\alpha = 0.05$) are used to test the significance of differences.

Workers in pleasant moods reported significantly higher UES-SF scores than those in unpleasant moods on conversational interfaces with an involvement style (Con+I) for executing Information Finding (pleasant: $\mu = 4.4$, $\sigma = 0.8$ vs. unpleasant: $\mu = 3.7$, $\sigma = 0.7$),

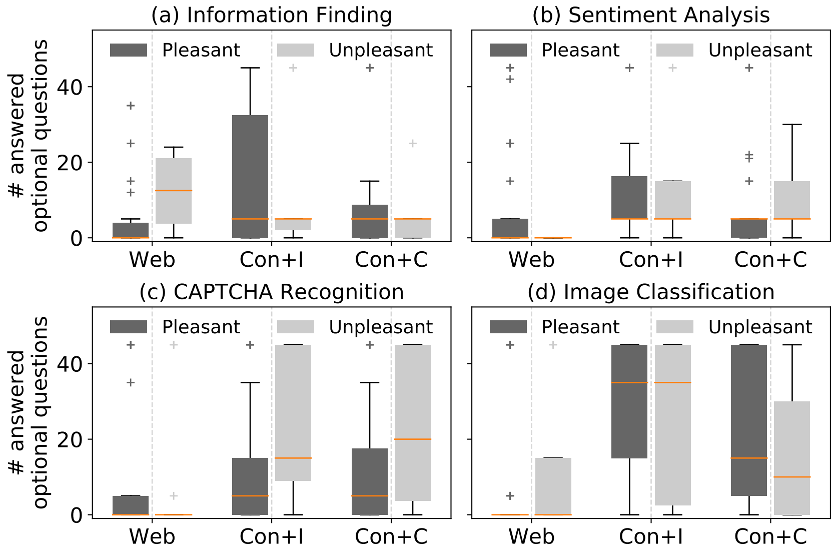


Figure 4.7: Boxplots showing the number of optional HITs completed by workers in different moods across different experimental conditions. Red lines in the boxplots represent the median value.

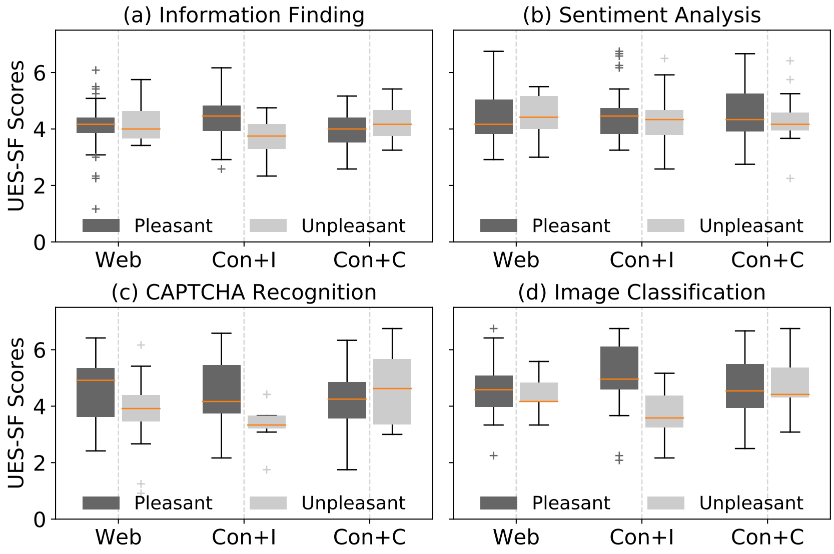


Figure 4.8: UES-SF scores across different experimental conditions and worker moods. Red lines in the boxplots indicate the median value.

CAPTCHA Recognition (pleasant: $\mu = 4.4$, $\sigma = 1.1$ vs. unpleasant: $\mu = 3.4$, $\sigma = 0.8$), and Image Classification (pleasant: $\mu = 5.1$, $\sigma = 1.1$ vs. unpleasant: $\mu = 3.8$, $\sigma = 0.8$) HITs (pleasant vs. unpleasant on IF Con+I, CR Con+I and IC Con+I, $p = 0.02$, $p = 0.014$ and $p = 0.0001$ respectively).

UES-SF scores of workers in unpleasant moods using conversational interfaces with a considerateness style (Con+C) were significantly higher than those using conversational interfaces with an involvement style (Con+I) in CAPTCHA Recognition (Con+I $\mu \pm \sigma = 3.4 \pm 0.8$ vs. Con+C $\mu \pm \sigma = 4.6 \pm 1.3$) and Image Classification (Con+I $\mu \pm \sigma = 3.8 \pm 0.8$ vs. Con+C $\mu \pm \sigma = 4.7 \pm 1.0$) HITs (unpleasant, Con+I vs. Con+C in CR and IC, $p = 0.036$ and $p = 0.0125$ respectively). The High-Involvement conversational interface ($\mu = 4.4$, $\sigma = 0.8$) corresponds to significantly higher UES-SF scores than the High-Considerateness conversational interface ($\mu = 3.9$, $\sigma = 0.7$) for workers in pleasant moods working on Information Finding HITs (pleasant, IF Con+I vs. IF Con+C, $p = 0.013$).

### Cognitive Task Load

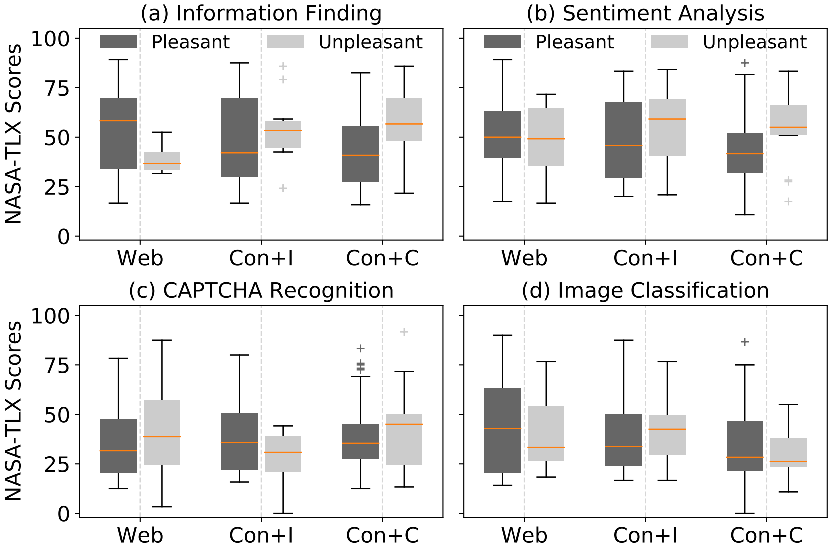


Figure 4.9: NASA-TLX scores different experimental conditions and worker moods. Red lines in the boxplots indicate the median value.

We also calculated the un-weighted NASA-TLX scores of all the workers participating in the crowdsourcing experiment. We use t-tests (two-tailed, $\alpha = 0.05$) to test the significance of differences between experimental conditions and worker moods.

Workers in pleasant moods reported significantly lower NASA-TLX scores than workers in unpleasant moods in conversational interfaces with a High-Considerateness style (Con+C) for Information Finding (pleasant $\mu \pm \sigma = 42.8 \pm 19.1$ vs. unpleasant $\mu \pm \sigma = 55.4 \pm 18.1$) and Sentiment Analysis (pleasant $\mu \pm \sigma = 43.3 \pm 17.2$ vs. unpleasant: $\mu \pm \sigma = 54.9 \pm 18.1$) HITs (pleasant vs. unpleasant on IF Con+C and SA Con+C, $p = 0.046$ and $p = 0.041$ respectively). Thus, workers in pleasant moods perceived lesser cognitive task load in these conditions. Moreover, workers in pleasant moods also perceived less cognitive load while executing the Information Finding HITs on the conversational interface with a

High-Considerateness style ($\mu = 42.8$, $\sigma = 19.1$), compared to the traditional web interface ($\mu = 53.5$, $\sigma = 21.1$) (pleasant, IF Con+C vs. IF Web, $p = 0.0200$).

### Design Implications

Our results clearly indicated that conversational crowdsourcing can improve worker retention in general irrespective of worker moods. Statistical tests revealed the fact that pleasant workers were more engaged than unpleasant workers in general. This calls for the development and adoption of conversational crowdsourcing, and for methods to induce pleasant moods prior to HIT execution. Our results also suggested that conversational interfaces with a High-Considerateness style exhibited the potential to improve engagement of workers in unpleasant moods, while a High-Involvement style exhibited a potential to further engage workers in pleasant moods, which implies the use of an appropriate conversational style can improve worker engagement in different moods. In terms of cognitive task load, our findings showed that workers in pleasant moods could perceive less task load than those in unpleasant moods while executing text-based HITs, especially when the conversational agent used a High-Considerateness style. These findings present opportunities for task routing based on worker moods and by leveraging different conversational styles.

### Limitations and Future Work

The mood distribution of workers is naturally unbalanced. However, the overall distribution of crowd worker moods are consistent with prior works that indicate a skew towards pleasant moods [67, 244]. It is however, not ethically sound to elicit unpleasant moods among workers to study the interaction between their moods and conversational styles of an agent.

Despite the measures we took to ensure the reliability of responses of workers, as with any research that involves human subjects using self-reporting tools, a threat to the validity of our findings is the veracity of the self-reported moods of workers. The future work could focus on the design of short and reliable measures of worker moods.

## 4.2 Self-Identification with Worker Avatars

To increase participant engagement and satisfaction, the use of video games or employing conversational interfaces, are two methods that have received attention in recent HCI literature. Relevant work in the field of games research has shown that self-identification with avatars can be effective in improving players' enjoyment and satisfaction [220, 14]. The contexts of games and crowd work are underlined by the need to motivate and engage participants, yet the potential of using worker avatars to promote identification and improve worker satisfaction in microtask crowdsourcing has remained unexplored. This is important to investigate, since using worker avatars and assigning avatars characteristics or personality traits can increase identification [220, 146]. Avatar identification has been studied from three perspectives — similarity identification, embodied identification, and wishful identification. Similarity identification refers to the identification related to the similarity between the avatar and the user; embodied identification refers to the identification of the feeling whether (and to what extent) the user is inside the avatar; and wishful identification represents the identification of avatar characteristics that the user would like to have. Prior works have shown that avatar appearance and characteristics can affect similarity and wishful identification respectively [96, 97], whereas embodied identification demands more avatar operations and interactions, which is very common in video games but not essential in crowdsourcing.

To operationalize similarity identification and wishful identification, in this study we support workers in (a) building their own representations by customizing the appearance of their avatars, and (b) characterizing their avatars before they begin task execution, by selecting one out of three desirable worker characterizations drawn from related literature (*diligent worker, competent worker, balanced worker*) [109, 68]. Since the influence of worker avatars in crowd work has remained unexplored, we know little about their impact on both traditional task interfaces as well as conversational crowdsourcing. We thereby delve into this comparison through our work. In this section, we address the following research questions:

**RQ4.3:** How do worker avatars affect worker experience and quality-related outcomes in traditional web and novel conversational interfaces?

**RQ4.4:** How can worker self-identification with their avatars be facilitated using avatar customization and worker characterization?

Addressing these RQs, we carried out a study to investigate the effectiveness of using worker avatars in microtask crowdsourcing. We explore whether using worker avatars and enabling avatar customization can reduce the perceived workload, increase the intrinsic motivation of workers, and improve quality-related outcomes.

We designed worker avatars, and studied the influence of avatar appearance customization and characterization of customized avatars. We implemented worker interfaces for microtask execution with avatar appearance customization and characterization selection affordances, based on both traditional web interfaces and novel conversational interfaces (conversational crowdsourcing). Experiments were performed with 360 crowd workers across six experimental conditions. Our results revealed that using avatar appearance and characterization customization had a significant impact on lowering the perceived task difficulty. In summary, our contributions are:

1. We found that combining worker avatars with conversational interfaces can effectively reduce the perceived cognitive task load and increase worker retention.
2. We found that workers who put more effort into avatar customization exhibited better performances with high accuracy.
3. Our analysis of the behavior and performance of workers indicates the occurrence of similarity and wishful avatar identification.

Our findings have important implications in terms of reducing perceived workload and improving the sense of success through task design in microtask crowdsourcing. As argued by prior work, this can be crucial to the sustainability of microwork marketplaces [117].

### Method: Using Worker Avatars for Crowdsourcing

To answer **RQ4.3** and understand how effective worker avatars are, based on the type of interface, we designed worker avatars in both traditional web interfaces and novel conversational interfaces (conversational crowdsourcing) for task execution. To answer **RQ4.4** and understand how avatar customization and characterization affects crowd work, we facilitated avatar customization and the selection of desired worker characterization across the web and conversational interfaces. To this end, we conducted a $3 \times 2$ between-subject study comparing three avatar conditions (without avatar, with avatar, with avatar and desirable characterizations) and two worker interfaces (Web and Chat), across two task types (Image Transcription and Information Finding). Addressing the **RQs**, we considered the following dependent variables – perceived workload, intrinsic motivation, and worker performance.

#### Avatar Design

***Avatar Appearance.*** We used an avatar library called *avataaars*[27] to create 2D avatars by combining a variety of attributes, i.e., clothes, hair, emotions, accessories, and colors. Figure 4.10 (a) shows the HTML-based panel for avatar appearance customization. In the avatar customization panel, we provided seven options for changing the avatar appearance: skin color, hair, facial hair, hair color, mood, accessories, and cloth color.

The avatar is initialized with three parameters — gender, skin color, and mood. With an aim to foster similarity identification during appearance customization, the information for these three parameters is acquired from workers using a short demographic survey before the actual task execution. Note that workers were free to customize their avatar as they wished to thereafter.

*1) Gender.* Hair and facial hair types are initialized according to workers' gender. If a worker identifies as **female**, the corresponding avatar will be initialized with longer hair and without facial hair. If a worker identifies as **male**, the corresponding avatar is initialized with shorter hair and random facial hair types (including no facial hair). If the gender type is **non-binary** or **others**, the hair and facial hair types are randomized. Note that the initialization of the hair and facial hair styles uses traditional gender stereotypes, to represent the gender difference and create the approximation of gender appearance. We are aware that the initialized avatar appearance might not be in-line with an individual's gender

[27] `https://avataaars.com`

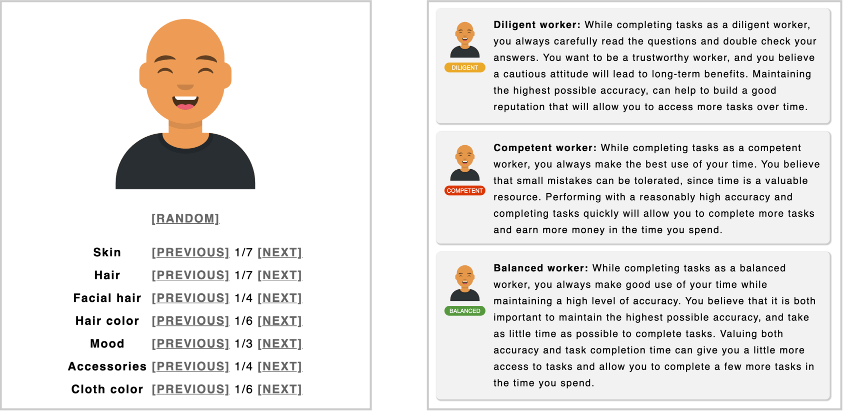
expression, therefore, all the workers have the freedom to change their hair and facial hair styles after initialization.

*2) Skin Color.* There are seven available skin colors for avatar initialization and customization, which are tanned, yellow, pale, light, brown, dark brown, and black.

*3) Mood.* Eyes, eyebrow, and mouth types are initialized according to workers' moods. Since we have shown the importance of worker moods in crowdsourcing (Section 4.1), we created a "mood" option by combining eyes, eyebrow, and mouth options. Note that in the original version of *avataaars*, the "mood" option does not exist – users need to customize moods by changing the emotion of eyes, eyebrow, and mouth. Using the mood option holistically instead of the individual attributes of eyes, eyebrow, and the mouth, we facilitate easy avatar appearance customization.

Apart from gender, skin, and mood, the accessories (types of glasses), and the color of the attire are randomly assigned for their initial avatar. After the avatar is initialized, workers have the freedom to change or randomize all previously mentioned options, as shown in Figure 4.10(a).

Considering that the most popular crowdsourcing marketplaces are Web-based, the avatars in our study are sketched on the Web-based interfaces using the vector format (SVG). Furthermore, the panel for avatar customization is purely based on HTML and JavaScript without any other dependencies. This makes the avatar customization very portable. Developers can easily deploy the avatar customization functionality to different Web applications with little overhead. The code repository for avatar appearance customization is shared publicly for the benefit of the community[28].



[RANDOM]

Skin [PREVIOUS] 1/7 [NEXT]
Hair [PREVIOUS] 1/7 [NEXT]
Facial hair [PREVIOUS] 1/4 [NEXT]
Hair color [PREVIOUS] 1/6 [NEXT]
Mood [PREVIOUS] 1/3 [NEXT]
Accessories [PREVIOUS] 1/4 [NEXT]
Cloth color [PREVIOUS] 1/6 [NEXT]

**Diligent worker:** While completing tasks as a diligent worker, you always carefully read the questions and double check your answers. You want to be a trustworthy worker, and you believe a cautious attitude will lead to long-term benefits. Maintaining the highest possible accuracy, can help to build a good reputation that will allow you to access more tasks over time.

**Competent worker:** While completing tasks as a competent worker, you always make the best use of your time. You believe that small mistakes can be tolerated, since time is a valuable resource. Performing with a reasonably high accuracy and completing tasks quickly will allow you to complete more tasks and earn more money in the time you spend.

**Balanced worker:** While completing tasks as a balanced worker, you always make good use of your time while maintaining a high level of accuracy. You believe that it is both important to maintain the highest possible accuracy, and take as little time as possible to complete tasks. Valuing both accuracy and task completion time can give you a little more access to tasks and allow you to complete a few more tasks in the time you spend.

(a) Appearance customization

(b) Characterization selection

Figure 4.10: The avatar (a) appearance customization and (b) characterization selection panels implemented based on HTML.

***Avatar Characterizations.*** According to self-discrepancy theory [92], the "actual self" represents one's self-concept, while the "ideal self" is the representation of characteristics that one would like to have. By customizing the appearance of their avatars, workers can build their actual-self representations. Combined with ideal characteristics (for example,

[28] `https://osf.io/x2bzp/?view_only=509b665ad7884e3180091228e68bb260`

competence or diligence) workers can create a model avatar. The objective is to explore whether an avatar that workers self-identify with can also have characteristics that workers aspire to (wishful identification).

We provide three ideal characterizations for workers to select, as can be seen in Figure 4.10 (b). We adopted these characterizations from previous work [109, 68]. Authors synthesized the characteristics of online crowd workers and grouped them into five main categories — Diligent, Competent, Spammers, Less-competent, and Sloppy workers. In the original work by Kazai et al. [109], Diligent workers were characterized by a high ratio of high-quality output, longer average time spent per task, and high label accuracy. In comparison, Competent workers produce many useful labels and obtain high accuracies, but work relatively faster. Sloppy workers were characterized by their low task completion time and concomitant low accuracy. Incompetent workers are characterized by their high task completion times and concomitant low accuracy. Spammers were characterized by their ulterior motives to complete tasks quickly and maximize their rewards (by gaming the tasks), resulting in very low accuracies. Spammers, Less-competent, and Sloppy workers are negative characterizations that workers would want to avoid on crowdsourcing platforms — being perceived as sloppy might have negative consequences for workers, e.g., privileges are revoked or completed tasks are rejected without pay [148]. Due to the impact of rejection on worker reputation and their future access to tasks, workers typically refrain from wilfully under-performing in tasks. Therefore, as shown in Figure 4.10 (b) we do not consider negative characterizations and used the characterizations of **Diligent** and **Competent** workers in our study as those characteristics that workers aspire to (wishful identification). Considering the accuracy and completion time factors during task execution, Diligent workers are defined to exhibit high accuracy, but correspond to long task execution time, while Competent workers exhibit reasonably high accuracy and short task completion time [68]. In this study, we adopted the definitions of Diligent workers and Competent workers. But on the user interfaces shown to workers, we added more details about the motivation behind a worker characterization (i.e. why a worker can be diligent/competent). For example, we show "maintaining the highest possible accuracy can help to build a good reputation that will allow you to access more tasks over time" in the description of Diligent worker. We also introduced a **Balanced** characterization, to represent an ideal worker type who maintains balanced levels of accuracy and task execution speed that workers may wish to possess or aspire to.

After customizing the appearance of their avatars, workers can select one out of these three characterizations for their avatars:

**Diligent worker:** *While completing tasks as a diligent worker, you always carefully read the questions and double-check your answers. You want to be a trustworthy worker, and you believe a cautious attitude will lead to long-term benefits. Maintaining the highest possible accuracy can help to build a good reputation that will allow you to access more tasks over time.*

**Competent worker:** *While completing tasks as a competent worker, you always make the best use of your time. You believe that small mistakes can be tolerated, since time is a valuable resource. Performing with reasonably high accuracy and completing tasks quickly will allow you to complete more tasks and earn more money in the time you spend.*

**Balanced worker:** *While completing tasks as a balanced worker, you always make good use of your time while maintaining a high level of accuracy. You believe that it is both important to maintain the highest possible accuracy, and take as little time as possible to complete tasks. Valuing both accuracy and task completion time can give you a little more access to tasks and allow you to complete a few more tasks in the time you spend.*

During the task execution, the selected and desired characterization is always displayed below the avatar in a `characterization label` (cf. Figure 4.10(b)).

***Avatar Conditions.*** To study whether avatar customization can affect crowdsourcing outcomes and workers' experience, workers were randomly assigned to three avatar conditions. In addition to a control condition without avatars, the avatar conditions were designed to operationalize self-identification with the avatars – to trigger similarity identification, and wishful identification. Each of these conditions is described below.

*1) Without avatar customization (hereafter referred to as* ***w/o avatar****).* This was set up to serve as a control condition, and allow us to compare workers' experience and performance to a condition unaffected by previously established motivational effects of using worker avatars.

*2) With avatar appearance customization (hereafter referred to as* ***w/ avatar****).* This condition was set up to investigate whether using avatar identification based on appearance customization as a means to facilitate similarity identification, can positively affect workers' experience and quality-related outcomes.

*3) With avatar appearance customization and worker characterization selection (hereinafter referred to as* ***w/ avatar+ch****).* This condition was set up to explore how characterization selection as a means to additionally facilitate wishful identification can affect worker performance and experience.

### Worker Interfaces

Addressing **RQ4.3**, we compare traditional web-based worker interfaces with novel conversational interfaces, not only to investigate the effect of using avatars in traditional microtask crowdsourcing, but also to study whether the use of avatars can have additional benefits for conversational interfaces.

Traditional web interfaces are the standard means for task execution on most crowdsourcing marketplaces such as Amazon's Mechanical Turk. We refer to the traditional web interfaces as **Web** in figures and tables henceforth. The traditional web interfaces are developed using HTML, CSS, and Javascript. On the web interface, all the essential elements of the task, such as task instructions, content of the microtasks, and corresponding input elements, are displayed on a single web page.

To investigate whether the use of avatars can further improve the effectiveness of conversational interfaces, we used TickTalkTurk for deploying text-based conversational crowdsourced microtasks on popular crowdsourcing platforms. The conversational interface deployed by TickTalkTurk is also built using HTML, CSS, and Javascript, and therefore compatible with most crowdsourcing platforms. We refer to the conversational interfaces as **Chat** in figures and tables henceforth. On the conversational interface, the task instructions, avatar customization, microtasks, and surveys are sent to workers via messages, from a gender-neutral conversational agent named "Andrea" with the profile image of a droid. Workers then reply to messages using a simple text field, or use the provided input elements (i.e., buttons, sliders) to respond to questions and tasks presented by the conversational agent.

Based on the task types being served, we provide three input types: *1) Single-selection*: this input type is used for workers to select one answer from multiple choices, which is implemented using `radio buttons` and `customized buttons` respectively on Web and Chat interfaces; *2) Free-text*: this input type is used for providing open-ended answers. Workers are required to input their answers via a `textarea` HTML element on the Web interface, or type their answers and send to the conversational agent as `messages` on the Chat interface; *3) Slider*: workers can move a handle to indicate a value on the slider. Both Web and Chat interfaces use HTML-based `slider` elements to provide input for some specific types of questions.

During task execution in the traditional Web interfaces, the customized avatar (either with or without the `characterization label`) is displayed on the left side of the input element that the worker is focusing on. We chose to position the avatar visibly to ensure that workers can always see the avatar and have opportunities to identify with their customized avatars, as shown in Figure 4.11 (b) and (c). Similarly, on the conversational Chat interface, the customized avatar is always displayed instead of the users' profile image, as shown in Figure 4.11 (e) and (f).

### Microtask Design

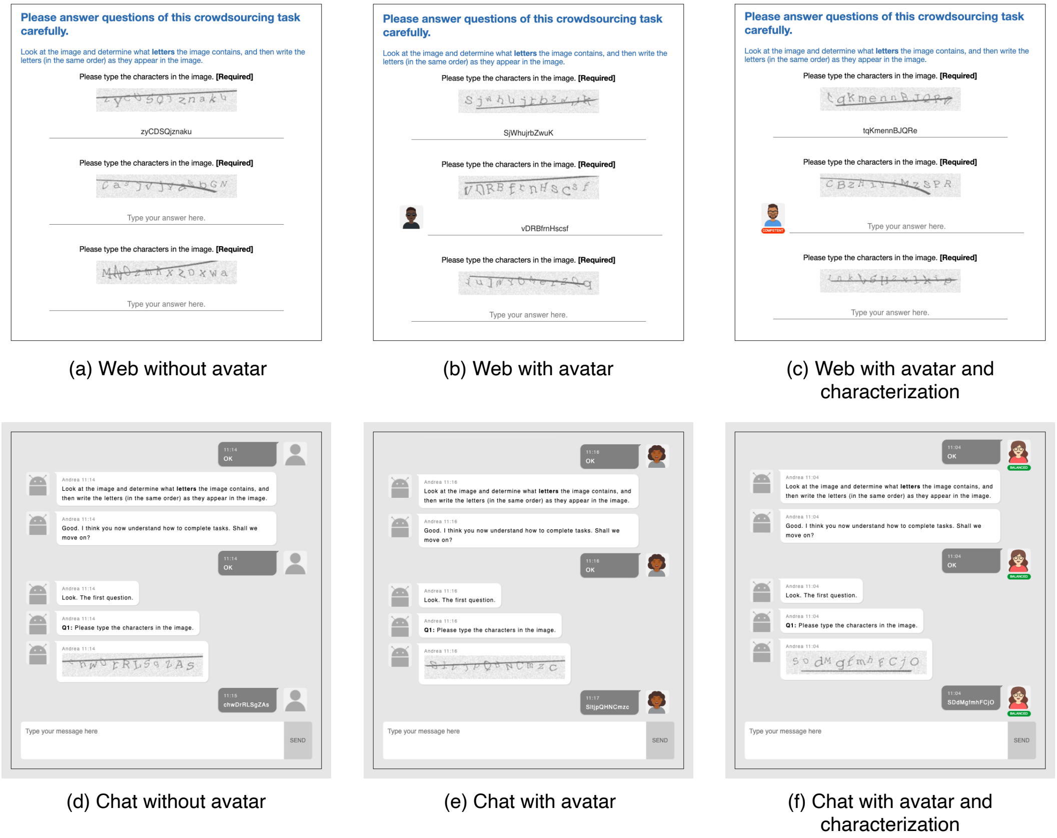
We chose the task types of Image Transcription and Information Finding to conduct our experiments, and investigated the impact of using avatar-related affordances on task performance across these two task types. These two task types are popularly crowdsourced [72, 51] and have been used in Section 3.2. Image Transcription tasks are relatively easy but can be highly monotonous. Information Finding tasks are relatively difficult, but workers can gain new knowledge while searching the web for relevant meta-data during task execution. In Chapter 2, we have shown that conversational interfaces can employ text input as an alternative to other types of input. For example, multiple-selection can be realized by asking users to type option labels/numbers. In this study, we only consider textual input as the input type for the tasks. A variety of input types, such as multiple-selection, sliders, or even bounding boxes, can be studied in the imminent future.

***Image Transcription.*** In these tasks, workers view the images randomly generated by Claptcha[29] and transcribe the text displayed in the images. By using Claptcha, the actual text in the image, the image size, and the strength of noise can be easily tuned. The images for transcription are automatically generated, containing 5 - 18 random, distorted English letters (upper case or lower case) with Gaussian white noise. Image Transcription microtasks need relatively less time and effort compared to the Information Finding tasks described below.

***Information Finding.*** In these tasks, workers are asked to find the middle name of a famous person by searching the web. We created a list of celebrities from different domains, including scientists, artists, politicians, musicians, and athletes. Workers are required to find the correct middle name according to given information, i.e., first and last names, with or without profession and active years in case there is ambiguity.

The celebrities in the list are selected to represent different level of complexity. For instance, finding the middle name of Alan Turing is not ambiguous, while the name of

[29] `https://github.com/kuszaj/claptcha`



**Please answer questions of this crowdsourcing task carefully.**

Look at the image and determine what **letters** the image contains, and then write the letters (in the same order) as they appear in the image.

Please type the characters in the image. **[Required]**

zyCDSQjznaku

Please type the characters in the image. **[Required]**

Type your answer here.

Please type the characters in the image. **[Required]**

Type your answer here.

(a) Web without avatar

**Please answer questions of this crowdsourcing task carefully.**

Look at the image and determine what **letters** the image contains, and then write the letters (in the same order) as they appear in the image.

Please type the characters in the image. **[Required]**

SjWhujrbZwuK

Please type the characters in the image. **[Required]**

vDRBfrnHscsf

Please type the characters in the image. **[Required]**

Type your answer here.

(b) Web with avatar

**Please answer questions of this crowdsourcing task carefully.**

Look at the image and determine what **letters** the image contains, and then write the letters (in the same order) as they appear in the image.

Please type the characters in the image. **[Required]**

tqKmennBJQRe

Please type the characters in the image. **[Required]**

Type your answer here.

Please type the characters in the image. **[Required]**

Type your answer here.

(c) Web with avatar and characterization

11:14
OK

Andrea 11:14
Look at the image and determine what **letters** the image contains, and then write the letters (in the same order) as they appear in the image.

Andrea 11:14
Good. I think you now understand how to complete tasks. Shall we move on?

11:14
OK

Andrea 11:14
Look. The first question.

Andrea 11:14
**Q1:** Please type the characters in the image.

Andrea 11:14

11:15
chwDrRLSgZAs

Type your message here

(d) Chat without avatar

11:16
OK

Andrea 11:16
Look at the image and determine what **letters** the image contains, and then write the letters (in the same order) as they appear in the image.

Andrea 11:16
Good. I think you now understand how to complete tasks. Shall we move on?

11:16
OK

Andrea 11:16
Look. The first question.

Andrea 11:16
**Q1:** Please type the characters in the image.

Andrea 11:16

11:17
SltjpQHNCmzc

Type your message here

(e) Chat with avatar

11:04
OK

Andrea 11:04
Look at the image and determine what **letters** the image contains, and then write the letters (in the same order) as they appear in the image.

Andrea 11:04
Good. I think you now understand how to complete tasks. Shall we move on?

11:04
OK

Andrea 11:04
Look. The first question.

Andrea 11:04
**Q1:** Please type the characters in the image.

Andrea 11:04

11:04
SDdMgfmhFCjO

Type your message here

(f) Chat with avatar and characterization

Figure 4.11: Worker interfaces for microtask crowdsourcing. (a), (b), and (c) represent traditional web worker interfaces (Web). (d), (e), and (f) represent novel conversational worker interfaces (Chat).

computer scientist Michael Jordan will also show results for the famous basketball player Michael Jordan. Compared to Image Transcription, each Information Finding microtask needs more time, but workers have an opportunity to gain new knowledge (e.g. to learn about more famous people and some potentially interesting facts) while completing the tasks.

### Measures

We use a variety of previously validated measures to understand workers' experience and performance. Self-reported surveys are used to measure the perceived workload of workers and their intrinsic motivation during task execution, while the worker performance is measured using accuracy in tasks and worker retention. In addition, we also analyze workers' behavior while customizing avatars and selecting characterizations.

***Perceived Workload.*** We use NASA's Task Load IndeX (NASA-TLX) questionnaire[30]

[30] https://humansystems.arc.nasa.gov/groups/TLX/

to measure workers' perceived workload. The NASA-TLX questionnaire evaluates worker's cognitive workload while completing tasks on six dimensions — Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Mental Demand and Physical Demand can measure how mentally or physically demanding the crowdsourcing task was. Temporal Demand can be interpreted as how hurried or rushed the pace of task execution was. Performance and Effort represent how successful the performance was and how hard the task was respectively, while accomplishing the task. Finally, Frustration indicates how stressed and annoyed the workers felt during task execution. Workers are required to indicate their feelings on each dimension using a slider ranging from 0 to 20. The TLX scores are later scaled to 0 to 100. The lower the TLX score is, the less mental demand, less physical demand, less temporal demand, more successful performance, less effort, and less frustration are perceived by the worker.

***Intrinsic Motivation.*** We use the Intrinsic Motivation Inventory (IMI) [145] to measure worker's intrinsic motivation to understand whether workers enjoy using the avatars, and thus how motivated during task execution they are. IMI has been widely used to assess play experience, and prior work has shown that self-identification with avatars can increase the intrinsic motivation of players [14, 84].

To reduce the workload for workers, we use a subset of the IMI covering the two most relevant dimensions — Interest-Enjoyment and Effort-Importance — consisting of 9 questions. Each question is answered by expressing agreement to statements on a 7-point Likert-scale from *1: strongly agree* to *7: strongly disagree.* The answers of the questions in IMI are provided using customized buttons and radio buttons on conversational interfaces and web interfaces respectively.

***Worker Accuracy.*** We use the percentage of correctly answered microtasks to measure worker accuracy. Specifically, in Information Finding tasks, a microtask is considered as correctly answered if and only if the answer provided by the worker contains the true middle name of the corresponding famous person, e.g., Irwin for Michael Jordan (computer scientist). In Image Transcription tasks, to maintain a reasonable task difficulty level, we added relatively strong artificial noises (Gaussian white noises, $0.1 \leq \sigma \leq 0.5$) and distortions into the images using Claptcha. This results in some completely illegible letters (roughly around 20% on manual inspection by the authors). Therefore, we use one of the most common string similarity metrics - the Levenshtein distance to measure the difference between the answer and the expected value [131]. In this work, we thereby tolerate 20% of mismatches. Thus, the answer for an image transcription microtask is considered to be correct if and only if the Levenshtein similarity ratio between the answer provided by workers and the expected value is greater than 80%. The Levenshtein similarity ratio is calculated as:

$$Levenshtein\ similarity\ ratio = \frac{|a| + |b| - lev(a,b)}{|a| + |b|}, \tag{4.1}$$

where $|a|$ and $|b|$ are the lengths of answer $a$ and the expected value $b$ respectively, while $lev(a,b)$ is the Levenshtein distance between the answer $a$ and the expected value $b$ (case insensitive). When the answer is identical to the expected value, the Levenshtein similarity ratio equals to 1. Furthermore, all spaces are stripped before calculating the Levenshtein similarity ratio.

***Worker Retention.*** We use the number of answered optional microtasks to measure worker retention. For each worker, there are at most 50 available microtasks (including

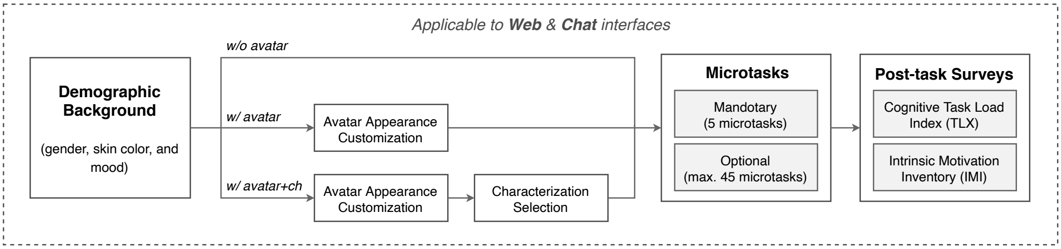

Applicable to Web & Chat interfaces
Demographic Background
(gender, skin color, and mood)
w/o avatar
w/ avatar
w/ avatar+ch
Avatar Appearance Customization
Avatar Appearance Customization
Characterization Selection
Microtasks
Mandotary (5 microtasks)
Optional (max. 45 microtasks)
Post-task Surveys
Cognitive Task Load Index (TLX)
Intrinsic Motivation Inventory (IMI)

Figure 4.12: Experimental procedure with three avatar conditions (*w/o avatar*, *w/ avatar*, and *w/ avatar+ch*).

mandatory microtasks and optional microtasks). As described earlier, workers first have to answer 5 mandatory microtasks, ensuring that we collect sufficient data for analyzing worker performance in terms of their accuracy and execution time. Workers cannot submit the answers if the 5 mandatory microtasks are not completed. After that, workers can complete as many of the 45 optional microtasks to follow as they wish.

## Experiments

In this study, we carried out experiments and recruited participants based on the Amazon's Mechanical Turk (MTurk) crowdsourcing platform. The study is approved by the Human Research Ethics Committee of TU Delft.

### Experimental Conditions

We conducted a 3 × 2 between-subject study across three avatar conditions (w/o avatar, w/ avatar, and w/ avatar+ch) and two worker interfaces (Web and Chat), resulting in six experimental conditions referred to as **Web w/o avatar**, **Web w/ avatar**, **Web w/ avatar+ch**, **Chat w/o avatar**, **Chat w/ avatar**, and **Chat w/ avatar+ch** to analyze worker experience and performance. With respect to the perceived workload of workers, their intrinsic motivation, and quality-related outcomes, we carried out analyses across two task types — Image Transcription and Information Finding.

### Procedure

The experiment is performed following the procedure displayed in Figure 4.12. Workers are required to first answer a few questions about their backgrounds. Before executing the crowdsourcing microtasks, workers in avatar-related experimental conditions will be guided through avatar customization. After executing five mandatory microtasks, workers can complete as many of the 45 optional microtasks as they wish to. Finally, workers are asked to complete two post-task surveys corresponding to their perceived workload and intrinsic motivation respectively. The details of the experimental procedure are explained below.

***Demographic Background.*** The objectives of asking demographic background questions are 1) to understand the demographic distribution of the workers, and 2) to initialize the avatar appearance according to workers' background. During this step, we ask workers three questions about gender, skin color, and mood respectively. There are four available gender options (non-binary, female, male, and others), seven available skin colors (tanned, yellow, pale, light, brown, dark brown, and black), and nine types of moods in three main categories (pleasant, unpleasant, and neutral). The instrument for measuring mood is Pick-A-Mood [48], which has been used in Section 4.1.

***Avatar Customization.*** The objective of avatar appearance customization and avatar characterization selection is to give workers an opportunity to finalize the desired appearance and characterization of their avatars, based on the initial avatar generated using the demographic background as a starting point. Depending on different experimental treatments, workers could either customize the appearance of their avatars, customize the appearance of their avatar and select a characterization for their avatar, or in case of the control condition – do neither.

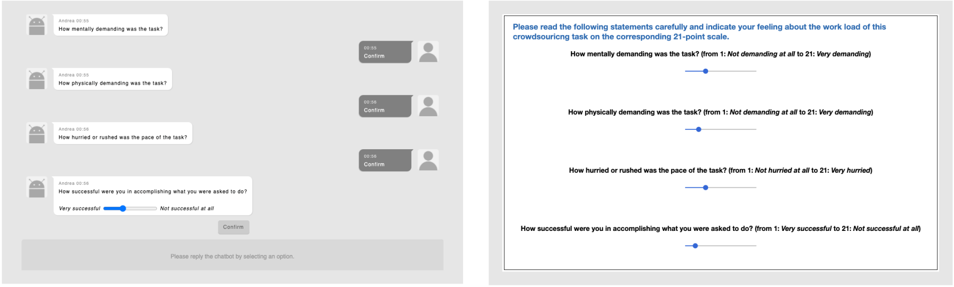
Workers assigned to the **w/o avatar** condition are directly asked to complete the microtasks (5 mandatory, 45 optional) after responding to the demographic background questions. While completing the microtasks in this condition, workers do not have a corresponding avatar, as shown in Figure 4.11 (a) and (d).

Workers in the **w/ avatar** condition have an opportunity to customize the visual appearance of their avatars. On completing the customization of their avatar's appearance, workers are asked to complete the microtasks (5 mandatory, 45 optional). Thus, a customized avatar is displayed throughout task execution, as shown in Figure 4.11 (b) and (e).

In the **w/ avatar+ch** condition, workers are required to do proceed through avatar appearance customization and avatar characterization selection, before they can complete the microtasks (5 mandatory, 45 optional). Therefore, the customized avatars are displayed while workers complete the microtasks, along with a `characterization label` below the avatars in each case, as shown in Figure 4.11 (c) and (f).

***Microtasks.*** During this step, workers are asked to complete actual microtasks. Each worker has to complete 5 mandatory microtasks. After completing 5 mandatory microtasks, workers can choose either to continue or stop task execution. We consider 45 optional microtasks that workers can complete to allow us to quantify worker retention based on the extent to which workers are willing to complete the available tasks. In the conversational interface (Chat) condition, the conversational agent, Andrea, asks workers whether they want to continue task execution or not, and then workers can indicate their decisions (yes or no) by clicking on customized buttons. On the traditional web interface (Web), workers can click a button stating "I want to answer more questions" to continue or directly end the task and continue with the post-task surveys. If a worker chooses to continue, they can complete as many of the 45 optional microtasks as they wish. Each time workers decide to continue, they are expected to complete another 10 optional microtasks until they ask to stop or continue to complete all the optional microtasks.

***Post-task Surveys.*** After completing the microtasks, workers are asked to complete two questionnaires. The first survey is the NASA-TLX for measuring workers' perceived workload. On both conversational interface (Chat) and traditional web interface (Web), workers should provide their answers using `slider` elements. In this study, workers in Chat conditions use the Chat interface and workers in Web conditions use the Web interface to complete post-task surveys. We did not redirect the workers in Chat conditions back to traditional web-based post-task surveys. This was motivated by prior work which has shown that dramatically changing UIs may affect users' mental models [191]. The second survey is a subset of the Intrinsic Motivation Inventory (IMI) for measuring workers' enjoyment and effort exerted during task execution.



Andrea 00:55
How mentally demanding was the task?

00:55
Confirm

Andrea 00:55
How physically demanding was the task?

00:56
Confirm

Andrea 00:56
How hurried or rushed was the pace of the task?

00:56
Confirm

Andrea 00:56
How successful were you in accomplishing what you were asked to do?

*Very successful* *Not successful at all*

Please reply the chatbot by selecting an option.

(a) the Chat interface for the post-task survey

Please read the following statements carefully and indicate your feeling about the work load of this crowdsouricng task on the corresponding 21-point scale.

**How mentally demanding was the task? (from 1: *Not demanding at all* to 21: *Very demanding*)**

**How physically demanding was the task? (from 1: *Not demanding at all* to 21: *Very demanding*)**

**How hurried or rushed was the pace of the task? (from 1: *Not hurried at all* to 21: *Very hurried*)**

**How successful were you in accomplishing what you were asked to do? (from 1: *Very successful* to 21: *Not successful at all*)**

(b) the Web interface for the post-task survey

Figure 4.13: User interfaces for workers to complete the post-task survey.

The post-task survey is implemented on both Chat and Web interfaces. Survey questions on both interfaces are exactly the same as the original metrics. A screenshot of survey interfaces is shown in Figure 4.13. Workers are expected to input the answer by using sliders/customized buttons in conversational interfaces, and sliders/radio buttons in traditional web interfaces respectively.

### Cost and Quality Control

We recruited participants from the Amazon's Mechanical Turk (MTurk) crowdsourcing platform. We set up 3 avatar conditions (**w/o avatar**, **w/ avatar**, and **w/ avatar+ch**) and 2 interfaces (**Web**, and **Chat**), resulting in $3 \times 2 = 6$ experimental conditions. For each condition, we published 60 Human Intelligence Tasks (HITs), and each HIT is completed by a unique worker following the between-subjects experimental procedure.

In order to avoid learning biases, each worker could complete only a single HIT throughout our entire experiment. To ensure this, we stored each worker's unique MTurk `WorkerID`. If a WorkerID was already recorded in our database, the task content was not rendered, and the corresponding worker was kindly informed to exit the HIT. In total, we recruited 60 unique workers per condition. In each condition, we randomly distributed the 60 workers into the two task types evenly (30 unique workers per task type). Thus, $60 \times 6 = 360$ unique workers participated in our experiment. To further ensure reliable participation, we used a qualification type provided by MTurk — each worker's overall HIT acceptance rate had to be greater than 95%. In addition, a worker who has one of the following behaviors is regarded as a malicious worker [73]: 1) accuracy is 0 and entering the same answer for all the questions; 2) accuracy is 0 and always entering meaningless random strings (not words). Therefore, we manually inspected the crowdsourced results and excluded 8 workers who exhibited obvious unreliable behavior. The excluded workers were not replaced in this study since they only account 2% of the total size..

Each worker was paid USD$1.5 for participating in our study and completing the surveys. Since avatar customization takes a very short amount of time in the context of the whole study, we paid workers for this time across all conditions irrespective of whether or not a worker customized the avatar. We payed a bonus of USD$0.02 per optional Image Transcription microtask or USD$0.05 per optional Information Finding microtask. Based on the average task execution time, including answering background questions and post-task surveys, the average hourly wage that workers received was nearly USD$11.50 (well above the federal minimum wage of USD$7.25 per hour).

## Results

### Demographic Distribution

Of all the 352 workers (8 were manually excluded) who participated in our experiment, 64% of workers (225) reported that they were male, while 36% of workers (125) reported that they were female. Two workers (less than 1%) identified as non-binary. As for skin colors, 42% of workers (149) indicated light skin; 19% (66) and 17% (61) of workers indicated brown and pale skin respectively; 29, 25, 14, and 8 workers indicated their skin color as black, tanned, yellow, or dark brown respectively. As for worker moods, most workers (82%, 287 workers) were in a pleasant mood, while 13% (46) of workers were in an unpleasant mood. The remaining 19 workers were in a neutral mood.

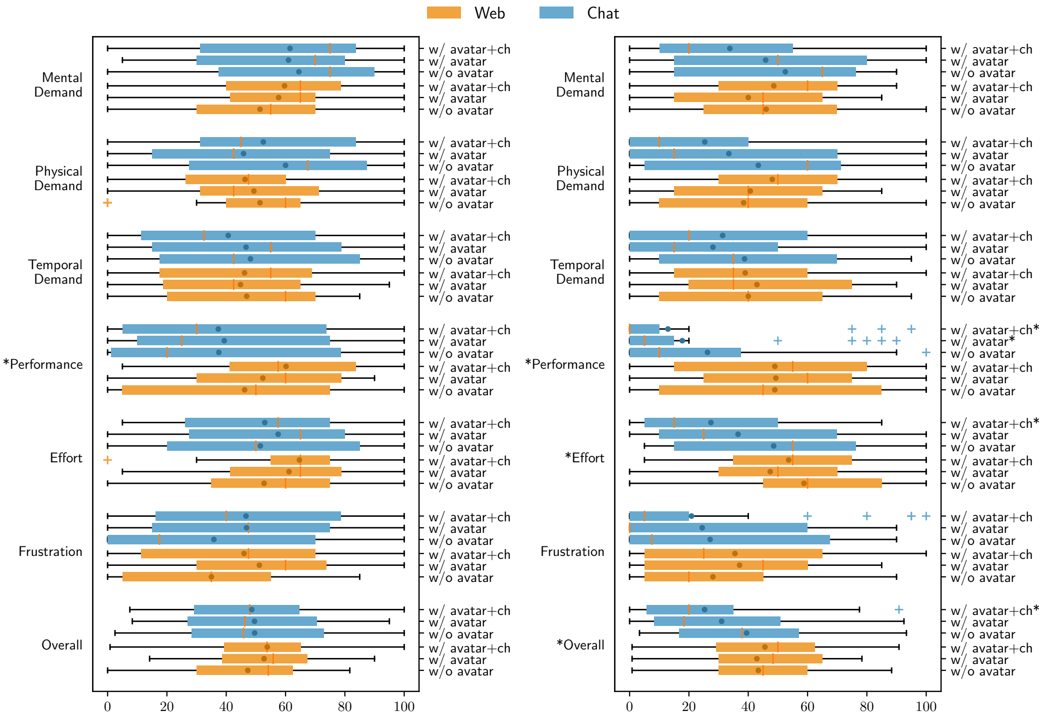
### Perceived Workload

The mean TLX scores for all six dimensions are illustrated in Figure 4.14. According to normality tests, the TLX score distributions come from a normal distribution (the average skewness is 0.2, the average kurtosis is -2.0, the average Shapiro-Wilk statistic score W is 0.88). To see if significant differences exist across the three avatar conditions (w/o avatar, w/ avatar, and w/ avatar+ch) and two interfaces (Chat and Web), we conducted two-way factorial multivariate ANOVA tests ($\alpha = 0.05$, Type I), with the null hypothesis that the mean value is the same across all six conditions (Web w/o avatar, Web w/ avatar, Web w/ avatar+ch, Chat w/o avatar, Chat w/ avatar, and Chat w/ avatar+ch). The results of the tests are shown in Table 4.2. For the Image Transcription tasks, we found that worker interfaces have a significant effect on the Performance dimension, showing that a conversational interface can significantly improve the sense of success with respect to performance. For the Information Finding tasks, we found that worker interfaces have significant effects on Performance, Effort, and the overall TLX score, suggesting that a conversational interface can reduce the perceived workload of workers. Furthermore, we found that conditions with avatars have a significant effect on the dimension of Effort, showing that avatar customization, either on Web or Chat interface, can significantly reduce the perceived task difficulty. We also observe a weak effect (not significant, $p = 0.067$) of the interaction of worker interfaces and avatar conditions in Physical Demand dimension.

Considering the web interface without avatar related affordances (which is the most commonly used interface in crowdsourcing tasks) as a baseline condition, we compared each other condition with the baseline (Web w/o avatar) using Bonferroni corrected independent t-tests (before correction $\alpha = 0.05$). All significant differences were found corresponding to Information Finding tasks.

In terms of self-reported Performance scores, we found the conversational interface with the two avatar customization conditions (Chat w/ avatar and Chat w/ avatar+ch) correspond to significantly better (lower) scores compared with the baseline ($p < 0.001$, Cohen's $d > 0.96$ for the two conditions). The conversational interface without avatar (Chat w/o avatar) could possibly lead to lower scores ($p = 0.02$ and Cohen's $d = 0.64$, not significant after Bonferroni correction).

Furthermore, the workers using conversational interfaces in the avatar appearance customization and characterization selection condition (Chat w/ avatar+ch) reported signifi-



(a) TLX scores of Image Transcription

(b) TLX scores of Information Finding

Figure 4.14: Boxplots of self-reported TLX scores corresponding to (a) *Image Transcription* and (b) *Information Finding* tasks across six dimensions. Dark points represent mean values and red lines (|) represent medians. The lower the TLX score is, the less mental demand, less physical demand, less temporal demand, more successful performance, less effort, and less frustration are perceived by workers respectively. Note that the asterisk (*) on a dimension indicates a statistically significant difference between conditions resulting from an ANOVA test; the asterisk (*) on a condition indicates a statistically significant difference in comparison with the baseline condition (web w/o avatar).

cantly lower Effort and overall TLX score in Information Finding tasks, compared to the baseline ($p < 0.006$, Cohen's $d > 0.77$). Workers in the condition with only avatar appearance customization (Chat w/ avatar) also reported lower Effort and overall TLX score ($p = 0.01$, $d = 0.71$ and $p = 0.03$, $d = 0.60$, $p < 0.05$ but not significant after Bonferroni correction).

To interpret our data beyond $p$-values and better understand effect sizes in terms of the overall TLX scores, we leverage estimation plots [94], as shown in Figure 4.15 (the estimation plots of other TLX dimensions can be found in the companion webpage[31]). Jitter plots show all the overall TLX scores, and how they distribute, across experimental conditions. Here, we use the baseline condition – the Web interface without worker avatars (the state of the art), as a control group in the plots, to make comparison with all the other experimental

[31] https://osf.io/x2bzp/?view_only=509b665ad7884e3180091228e68bb260

Table 4.2: Results of two-way multivariate ANOVA tests (Type I) for TLX dimensions and two-way ANOVA tests (Type I) for overall TLX scores.

| ***TLX Dimension*** | ***Factor*** | **Image Transcription** | | | **Information Finding** | | |
|---|---|---|---|---|---|---|---|
| | | Df | F-Value | Pr(>F) | Df | F-Value | Pr(>F) |
| | *Worker Interface (W)* | 1 | 1.818 | 0.179 | 1 | 0.039 | 0.844 |
| *Mental Demand* | *Avatar Condition (A)* | 2 | 0.111 | 0.895 | 2 | 1.047 | 0.353 |
| | *W × A* | 2 | 0.620 | 0.539 | 2 | 2.190 | 0.115 |
| | *Worker Interface (W)* | 1 | 0.641 | 0.424 | 1 | 3.067 | 0.082† |
| *Physical Demand* | *Avatar Condition (A)* | 2 | 1.082 | 0.341 | 2 | 0.291 | 0.748 |
| | *W × A* | 2 | 0.612 | 0.544 | 2 | 2.747 | 0.067† |
| | *Worker Interface (W)* | 1 | 0.026 | 0.872 | 1 | 2.883 | 0.091† |
| *Temporal Demand* | *Avatar Condition (A)* | 2 | 0.238 | 0.788 | 2 | 0.319 | 0.727 |
| | *W × A* | 2 | 0.233 | 0.793 | 2 | 0.699 | 0.499 |
| | *Worker Interface (W)* | 1 | 8.649 | 0.003* | 1 | 37.251 | 7.03e-9* |
| *Performance* | *Avatar Condition (A)* | 2 | 0.608 | 0.545 | 2 | 0.599 | 0.550 |
| | *W × A* | 2 | 0.679 | 0.508 | 2 | 0.622 | 0.538 |
| | *Worker Interface (W)* | 1 | 1.718 | 0.192 | 1 | 11.608 | 0.0008* |
| *Effort* | *Avatar Condition (A)* | 2 | 1.160 | 0.316 | 2 | 3.147 | 0.046* |
| | *W × A* | 2 | 0.541 | 0.583 | 2 | 1.258 | 0.287 |
| | *Worker Interface (W)* | 1 | 0.043 | 0.836 | 1 | 3.666 | 0.057† |
| *Frustration* | *Avatar Condition (A)* | 2 | 2.697 | 0.070† | 2 | 0.160 | 0.852 |
| | *W × A* | 2 | 0.113 | 0.893 | 2 | 0.744 | 0.477 |
| | *Worker Interface (W)* | 1 | 0.677 | 0.412 | 1 | 15.322 | 0.0001* |
| *Overall TLX* | *Avatar Condition (A)* | 2 | 0.226 | 0.798 | 2 | 0.971 | 0.381 |
| | *W × A* | 2 | 0.370 | 0.691 | 2 | 1.647 | 0.196 |

Note: † means $0.05 \leq p < 0.1$, and * means $p < 0.05$

conditions. The estimation plots also show the resampling distribution of the difference in means, representing the effect size. We found that the effect sizes in Image Transcription tasks were minor. However, it is still obvious that in jitter plots (swarm plots), the points corresponding to conversational interfaces tend to distribute below 50 (the middle point of TLX scale), while the points corresponding to traditional Web interfaces tend to distribute above 50. In terms of Information Finding, in comparison with the baseline (Web w/o avatar), the effect sizes of worker avatars on conversational interfaces (both Chat w/ avatar and Chat w/ avatar+ch) are large, showing a possible positive impact of the interaction effect of conversational interface and worker avatar on perceived workload.

***Summary.*** Our results suggest that — ***i)*** *The conversational interface generally corresponds to lower perceived workload compared to the Web interface, particularly in Information Finding tasks.* ***ii)*** *Worker avatars can reduce perceived task difficulty compared to the no-avatar condition in Information Finding tasks.* ***iii)*** *The conversational interface with the affordance of avatar appearance customization and avatar characterization selection (Chat w/ avatar+ch) can improve the workers' perceived success and difficulty while completing tasks.*

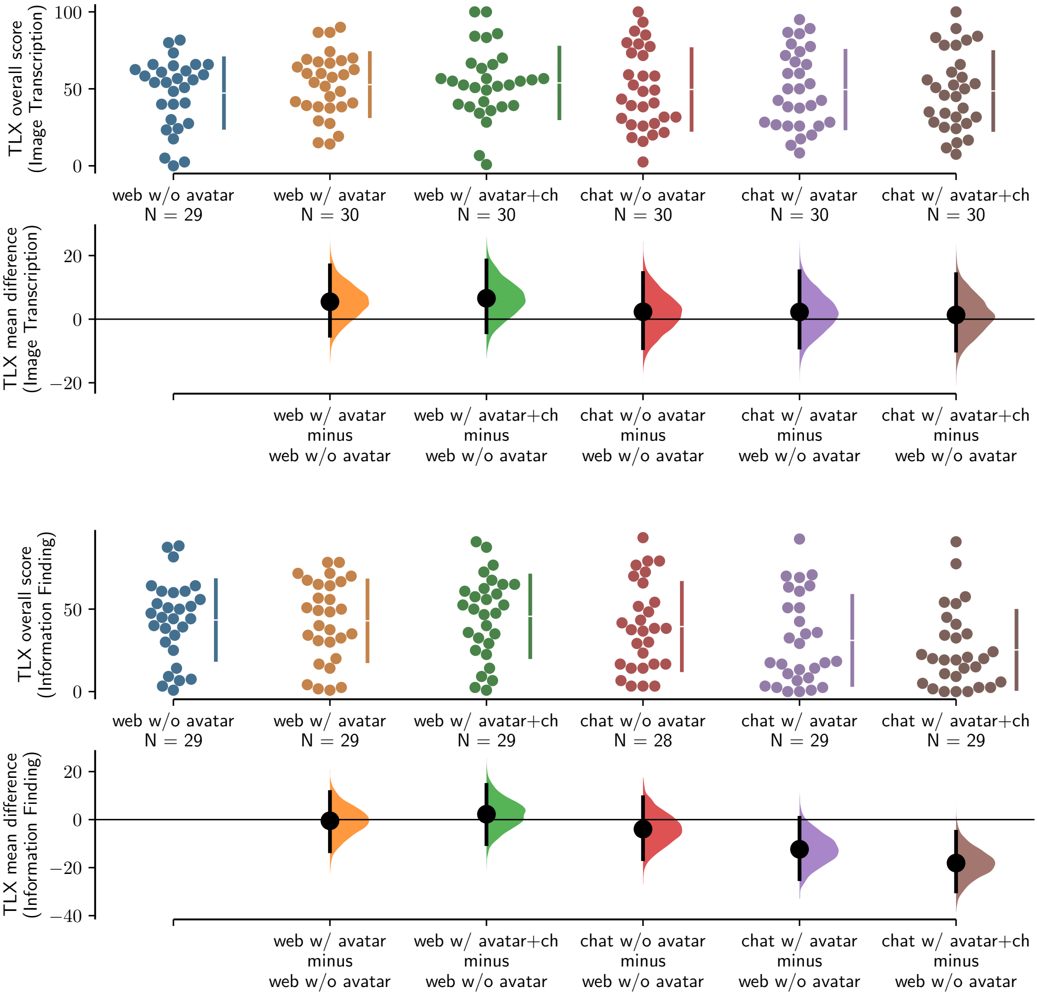
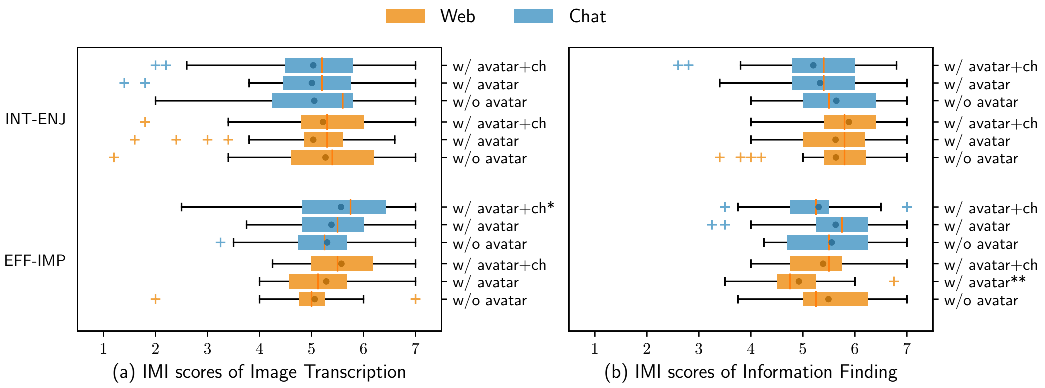


Figure 4.15: Estimation plots of TLX scores of Image Transcription and Information Finding tasks.

### Intrinsic Motivation

Figure 4.16 shows the IMI scores of different avatar conditions across two interfaces (Web and Chat) and two task types (Image Transcription and Information Finding), in two intrinsic motivation dimensions — Interest-Enjoyment and Effort-Importance respectively.

According to normality tests, IMI score samples (across the two task types, two IMI dimensions and six conditions) come from a normal distribution (the average skewness is -0.9, the average kurtosis is 0.3, the average Shapiro-Wilk statistic score W is 0.93). We thereby used independent t-tests with Bonferroni correction (before correction $\alpha = 0.05$) to test the null hypothesis that the IMI scores of experimental conditions come from the same distribution, compared to the baseline condition (Web w/o avatar).



(a) IMI scores of Image Transcription

(b) IMI scores of Information Finding

Figure 4.16: Boxplots of self-reported intrinsic motivation inventory score of *Image Transcription* tasks and *Information Finding* tasks in interest-enjoyment and effort-importance dimensions, where dark points represent mean values and red lines (|) represent medians. Note that ** indicates statistical significance with Bonferroni correction and * indicates $p < 0.05$ but not significant after Bonferroni correction).

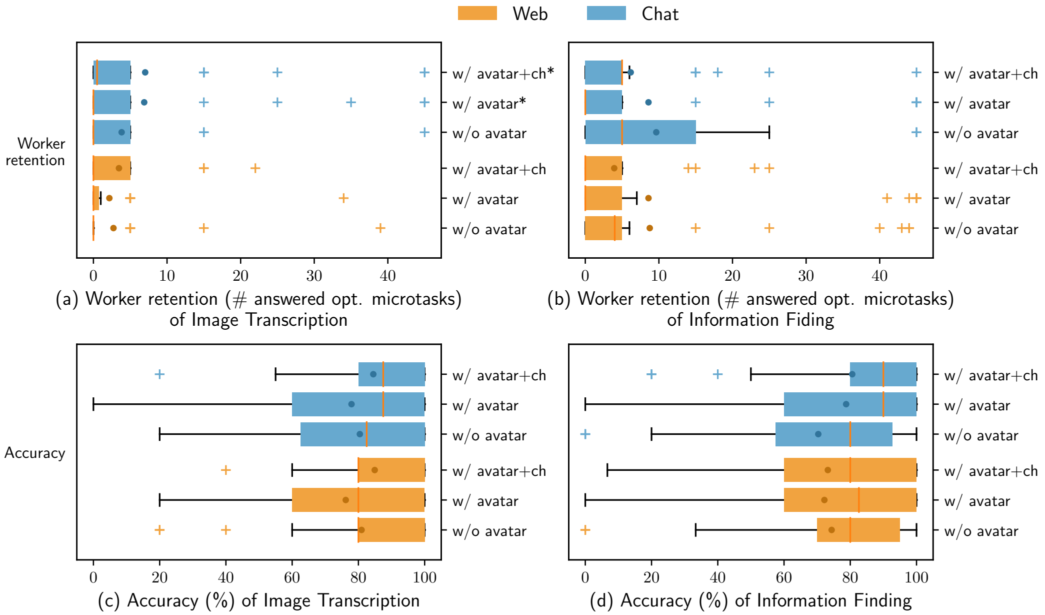
We did not find significant differences in the Interest-Enjoyment dimension. However, with respect to the Effort-Importance dimension that represents how important the task is, so that a worker needs to exert effort (note that Effort-Importance dimension in IMI is different from the Effort dimension in TLX which represents the perceived task difficulty), we found significant differences (after Bonferroni correction) corresponding to Information Finding tasks ($p = 0.007$, $d = 0.75$), where the Effort-Importance score of the Web w/ avatar condition is significantly lower than the baseline (Web w/o avatar). Thus, our results suggest that workers in the Web w/ avatar condition considered the Information Finding tasks to be relatively less important. Furthermore, in Image Transcription tasks, workers with avatar customization and characterization selection in the Chat interface (Chat w/ avatar+ch) reported higher EFF-IMP scores in comparison with the baseline (Web w/o avatar) with $p$-values equaling 0.025 ($d = 0.61$). However, this difference is not significant after Bonferroni correction is applied. It suggests that workers with avatar appearance customization and characterization selection may take the task more seriously and exert more effort in order to perform better.

***Summary:*** *Our findings suggest that avatar customization does not have a significant effect on worker intrinsic motivation, in either conversational interfaces or traditional web interfaces.*

### Objective Worker Performance

***Worker Retention.*** The results of worker retention, measured by the number of answered optional questions, are shown in Figure 4.17 (a) and (b). According to normality tests, the worker retention does not follow a normal distribution (the average skewness is 4.3, the average kurtosis is 3.1, the average Shapiro-Wilk statistic score is 0.57). Therefore, we used Mann-Whitney $U$ tests to find differences in worker retention across conditions measured by the number of answered optional microtasks. The results are in-line with our findings in Chapter 3. The conversational interfaces (Chat) were found to be more effective in retaining workers in both Image Transcription and Information Finding tasks, compared to Web interfaces ($p = 0.026$, CL effect size $f = 0.57$, and $p = 0.085$, CL effect size $f = 0.56$ respectively).

Particularly, in Image Transcription tasks, workers who used a conversational interface with avatars, either without or with characterization selection (Chat w/ avatar and Chat w/ avatar+ch, $p = 0.037/f = 0.62$, and $p = 0.018/f = 0.64$ respectively), completed more optional microtasks in comparison with the baseline condition — the Web interface without avatars.



(a) Worker retention (# answered opt. microtasks) of Image Transcription

(b) Worker retention (# answered opt. microtasks) of Information Fiding

(c) Accuracy (%) of Image Transcription

(d) Accuracy (%) of Information Finding

Figure 4.17: Boxplots of worker retention measured by the number of answered optional microtasks, and worker accuracy (%) measured by the percentage of correctly answered microtasks, corresponding to *Image Transcription* tasks and *Information Finding* tasks. Dark points represent mean values, red lines (|) represent medians, and (* represents significant difference in comparison with the baseline).

***Worker Accuracy.*** Results pertaining to worker accuracy are shown in Figure 4.17 (c) and (d). Aligned with the results from Section 3.1, we found no significant difference between experimental conditions across the two task types — Image Transcription ($p > 0.18$) and Information Finding ($p > 0.1$), according to Mann-Whitney $U$ tests (since worker accuracy does not come from normal distributions as per normality tests: the average skewness is -2.7; the average kurtosis is 1.2; the average Shapiro-Wilk statistic score is 0.81). However, as shown in Table 4.3, we found that the condition with avatar appearance customization and characterization selection (w/ avatar+ch) corresponds the highest worker accuracy in three out of four cases (Image Transcription on both Web and Chat interfaces, and Information Finding on Chat interface). Apart from our observation in the Information Finding tasks on the Web interface (where all three avatar conditions correspond to similar worker accuracy with only 1-2% differences), the mean values of worker accuracy of the avatar appearance customization and characterization selection condition (w/ avatar+ch) are 5%-13% higher than the baseline condition (w/o avatar).

Table 4.3: Worker accuracy (unit:%, $\mu \pm \sigma$) measured by the percetage of correctly answered microtasks, where the highest values among each interface are displayed in bold.

| | **Image Transcription** | | **Information Finding** | |
|---|---|---|---|---|
| ***Condition*** | Web | Chat | Web | Chat |
| *w/o avatar* | $80 \pm 21$ | $80 \pm 22$ | $\mathbf{74 \pm 27}$ | $70 \pm 31$ |
| *w/ avatar* | $76 \pm 22$ | $77 \pm 23$ | $72 \pm 32$ | $78 \pm 25$ |
| *w/ avatar+ch* | $\mathbf{84 \pm 15}$ | $\mathbf{84 \pm 17}$ | $73 \pm 30$ | $\mathbf{80 \pm 23}$ |

***Summary: i)*** *Our observation that the conversational interfaces can significantly improve worker retention is consistent with prior findings in HCI.* ***ii)*** *We found evidence that the use of worker avatars has a positive effect on worker retention.* ***iii)*** *The affordance of avatar customization with worker characterization selection shows an increasing trend in worker accuracy, although our results are inconclusive in this regard.*

### Avatar Appearance Customization and Characterization Selection

***Avatar Customization Time.*** In terms of the time spent on appearance customization, workers in the Web interface conditions (29.66 ± 31.76 seconds) spent slightly longer time on customizing avatars in comparison with workers in the Chat interface conditions (23.35 ± 31.39 seconds). To analyze the impact of customization time on worker performance, we split the workers into three groups according to the standard deviation ($\pm 0.5\sigma$) of avatar customization time, resulting in — a group of workers with short customization time (customization time $< \mu - 0.5\sigma$, less than 10.6 seconds), a group of workers with medium customization time ($\mu - 0.5\sigma \leq$ customization time $< \mu + 0.5\sigma$, 10.6-42.3 seconds), and a group of workers with long customization time (customization time $\geq \mu + 0.5\sigma$, longer than 42.3 seconds). As shown in Table 4.4, we found that the group of workers corresponding to a long customization time exhibit the highest worker accuracy ($83 \pm 22\%$) in comparison with the group of workers with short customization time (accuracy $= 77 \pm 27\%$, $p = 0.065$, CL effect size $f = 0.57$), and with medium customization time (accuracy $= 77 \pm 24\%$, $p = 0.021$, CL effect size $f = 0.60$) using Mann-Whitney $U$ tests. The results suggest that workers spending a longer time on avatar customization go on to perform with a higher accuracy in general. This may be explained by a greater level of self-identification through avatar customization which leads to an increased intrinsic motivation, as supported by our findings.

In terms of task execution time, we found the group of workers with long customization time spent significantly longer time on task execution (62.63 ± 70.13 seconds), compared to the group of workers with short customization time (execution time $= 42.09 \pm 36.69$ seconds, $p = 0.024$, Cohen's $d = 0.37$) or with medium customization time (execution time $= 43.01 \pm 34.77$ seconds, $p = 0.029$, Cohen's $d = 0.35$) using independent t-tests. This reveals that workers who spent more time in customizing their avatars also took longer to execute tasks.

Table 4.4: Worker accuracy (unit: %, $\mu \pm \sigma$) and execution time per microtask (unit: second, $\mu \pm \sigma$) of three groups divided according to avatar customization time.

| ***Avatar Customization Time***<br>*(seconds)* | | **Worker Accuracy**<br>(%) | **Task Execution Time**<br>(seconds per microtask) |
|---|---|---|---|
| *Short* | (*time* $< 10.6$) | $77 \pm 27$ | $42.09 \pm 36.69$ |
| *Medium* | ($10.6 \leq$ *time* $< 42.3$) | $77 \pm 24$ | $43.01 \pm 34.77$ |
| *Long* | (*time* $\geq 42.3$) | $83 \pm 22$ | $62.63 \pm 70.13$ |

For all the workers in avatar conditions, we found that on average, the avatar customization time ($26.48 \pm 31.73$ seconds) occupies 8.14% (±10.01%) of the total task execution time ($426.03 \pm 399.73$ seconds). From the perspective of the task requester, facilitating avatar customization may not appear to be a useful investment for short or less complex batches of tasks, considering the additional costs that requesters may incur in return for limited positive effects. However, for long, complex, or challenging tasks, facilitating avatar customization can warrant the reasonable overheads with an aim to effectively improve worker experience. We envision that in the future, avatar customization can be a feature that is supported by crowdsourcing platforms rather than by individual task requesters with an aim to foster a healthy work experience for crowd workers.

***Self-Identification with Worker Avatars.*** We explored the number of workers who actually changed the appearance of their initial avatar, that was generated with 3 parameters in accordance to their demographic backgrounds. We found that only 58 workers (24%) changed their skin colors for their avatars, while most of these workers (37 out of 58) just slightly tuned the skin color (for instance, change between Black and Dark Brown, Dark Brown and Brown, Brown and Light, or Light and Pale), suggesting that most workers were generally satisfied with their initialized avatars based on the demographic information they provided. We found that 26 workers (11%) changed the depicted moods for their avatars. 19 out of 26 workers changed their moods from either pleasant or unpleasant to neutral; only 4 workers changed to pleasant moods from unpleasant moods, while 3 workers did the reverse. As for accessories and clothing colors, we found that all the types of accessories are nearly equally distributed, and more workers chose Black for avatar's clothing while the least number of workers chose Yellow (50 Black and 29 Yellow, average = 40). Our findings pertaining to avatar customization indicate that workers generally cared about the appearance of their avatars, and this suggests the potential emergence of self-identification [14].

As for the results of characterization selection, of 118 workers (2 unreliable workers were excluded from 120) who were in the condition of avatar appearance customization and characterization selection (w/ avatar+ch on both Web and Chat), 33 workers selected the "Diligent" characterization for their avatar; 17 workers selected "Competent"; and 68 workers selected the "Balanced" characterization.

Diligent workers are described as workers who exhibit a high accuracy but a relatively slower task execution speed, and Competent workers are described as workers with a reasonably high accuracy but a faster task execution speed. As shown in Table 4.5, on exploring the mean worker accuracies of the workers who selected the "Competent" characterization, we found that they exhibited a higher accuracy than the workers who selected a "Diligent" characterization in both Image Transcription and Information Finding tasks (9% and 24%

Table 4.5: Worker accuracy (unit: %, $\mu \pm \sigma$) and execution time per microtask (unit: second, $\mu \pm \sigma$) of *Image Transcription* tasks and *Information Finding* tasks across *Diligent*, *Competent*, and *Balanced* characterizations.

| ***Measure*** | ***Characterization*** | **Image Transcription** | **Information Finding** |
|---|---|---|---|
| Worker accuracy | Diligent | $80 \pm 15$ (N = 19) | $68 \pm 36$ (N = 14) |
| | Competent | $87 \pm 8$ (N = 6) | $84 \pm 17$ (N = 11) |
| | Balanced | $87 \pm 17$ (N = 35) | $78 \pm 25$ (N = 33) |
| Execution time | Diligent | $33.93 \pm 19.25$ (N = 19) | $55.47 \pm 23.10$ (N = 14) |
| | Competent | $32.56 \pm 25.15$ (N = 6) | $59.44 \pm 46.29$ (N = 11) |
| | Balanced | $30.07 \pm 19.95$ (N = 35) | $65.20 \pm 55.53$ (N = 33) |

higher respectively). Interestingly, in terms of task execution time per microtask (speed), the workers who selected the "Diligent" characterization exhibited faster execution speeds in comparison with those workers who selected either the "Competent" or "Balanced" characterizations (7% and 15% faster respectively) in Information Finding tasks. This can be explained by workers' wishful identification with selected characterizations – workers were probably aware of their real characterizations (and shortcomings), therefore they may have chosen avatar characterizations which they aspired to. This is consistent with what has been observed in gaming research [14].

***Summary:*** ***i)*** *We found that workers who spent a long time on avatar customization exhibited a high accuracy.* ***ii)*** *Our analysis suggests that the appearance of a worker's avatar might represent their actual self, while the avatar's characterization might represent their ideal self. Additional experiments are required to further tease out the nuances of worker self-identification through avatar customization.*

### Discussion

Our study has shown that using worker avatars during task execution in general, can help workers perceive less difficulty during task execution. Using avatars in conversational interfaces can generally reduce the perceived workload of workers, increase intrinsic motivation, and improve user retention.

By analyzing the results about avatar appearance customization and characterization selection, we found that customization of the avatar's appearance facilitates similarity identification among workers, while the avatar characterization facilitates wishful identification among workers. Our results show that 58% of workers selected the "Balanced" characterization. Furthermore, we found that the performance of workers, to a large extent, does not follow the avatar characterization they selected. For example, 28% of workers who selected a Diligent characterization ended up performing with a relatively low accuracy, 14% of workers who selected a Competent characterization ended up exhibiting relatively long task execution times. This can potentially be explained by the emergence of wishful identification with avatar characterizations, and these workers falling short of their aspirations (i.e., to complete tasks with higher accuracies or lower task execution times respectively). Additional experiments are required to distill the extent to which avatar appearance customization, and avatar characterization in addition to appearance customization systematically facilitate similarity and wishful identification. Nevertheless, our findings bear evidence to support that avatar identification can be an effective tool for improving satisfaction and enjoyment during crowdsourcing task execution [220, 114].

### Using Avatars in Conversational Interfaces to Improve Crowdsourcing

Conversational interfaces for microtask crowdsourcing have emerged owing to the concomitant advantages of better engaging users. Our study has revealed that using avatars in conversational interfaces can ease the perceived workload of workers, by improving the sense of success in performance and reducing the perceived task difficulty while completing tasks. Specifically for tasks that are more challenging, we found that conversational interfaces with avatar appearance customization and avatar characterization selection can be effective in significantly reducing cognitive workload from the perspectives of performance (75% lower TLX score), and effort (53% lower TLX score), in comparison to the baseline of traditional web interfaces without worker avatars.

### The Value of Using Worker Avatars in Crowdsourcing

We note that conversational interfaces reduce the perceived cognitive load of workers in comparison to traditional web interfaces, corroborating recent findings. Critically reflecting on our collective findings, the added value of using worker avatars is less prominent in relatively easy tasks such as solving CAPTCHAs (Image Transcription), and along the dimensions of Mental Demand, Physical Demand, Temporal Demand, and Frustration. We found that using worker avatars in relatively more difficult tasks (Information Finding), led to a reduction in the perceived cognitive load of workers in conversational interfaces. Our findings hint that worker avatars can play a more significant role in tasks that are relatively more difficult, but include elements of learning (as indicated by the open-ended comments from workers). Another explanation for our null findings with respect to the impact of avatar customization in image transcription tasks can be that due to the relatively less amount of time required for task completion, workers do not meaningfully self-identify with their avatars. Future experiments can explore how self-identification with avatars is mediated by the task execution time.

## Design Implications

### Alleviating Perceived Workload

Our results showed that the conversational interface with the functionality of avatar appearance customization and characterization selection could effectively decrease cognitive workload. Particularly, the conversational interface with worker avatars could significantly make workers feel more successful, and perceive less difficulty, while completing tasks. This finding has important implications in future crowdsourcing task design. The mental state of crowd workers has become a major concern due to an increase in the number of workers who work full-time and earn their livings in crowdsourcing marketplaces, coupled with power asymmetry and other challenges workers typically face [104, 190, 83, 141]. Most state-of-the-art tools and approaches in the field of crowdsourcing are developed from the perspective of task requesters. Although human factors and worker-friendly interventions have been considered, the focus has largely been on improving the quality of outcomes [117, 110], rather than ensuring the wellbeing and the mental health of workers. The avatar customization framework we introduce in this work is developed completely based on HTML/CSS/Javascript, and is designed to be compatible and portable. Using the avatar framework we designed and made publicly available, there is very little overhead involved in integrating the use of worker avatars in microtask crowdsourcing platforms – task requesters can readily integrate avatar customization into their tasks. In exchange for the small overhead of integrating avatar customization, task requesters can reap worthy benefits of reducing the perceived task difficulty among workers and increase their sense of success.

#### Facilitating Avatar Identification

Avatar identification in crowdsourcing can be interpreted as the resonance and identification of crowd workers with an avatar that represents them. Avatar identification has been shown to be useful for fostering intrinsic motivation, increasing satisfaction and entertainment, and improving preventive health outcomes [14, 220, 114]. Our results pertaining to avatar appearance customization and characterization selection imply that similarity identification and wishful identification can be facilitated among workers in microtask crowdsourcing to reap similar rewards.

#### Strengthening the Requester-Worker Relationship

A strong and healthy relationship between task requesters and workers is crucial to all relevant stakeholders in the crowdsourcing paradigm. Maintaining a good relationship can assist task requesters in building their reputation and attracting more workers of high quality. For crowd workers, a good relationship can help them maintain credible profiles, increase their access to more work in the crowdsourcing marketplaces, maximize their earnings, and reduce the emotional toll and frustration that can result from mistrust and rejection [148]. By alleviating the perceived workload and improving the sense of success among workers through the use of avatars and conversational interfaces, there lies a great potential to further foster healthy requester-worker relationships. This bears a useful implication on ensuring the sustainability of crowdsourcing marketplaces.

### Limitations and Future Work

Previous studies about motivations in gaming systems have shown the effectiveness of fostering intrinsic motivation of the player by using avatar customization [14]. However, critically reflecting on our findings in microtask crowdsourcing, we found that avatar appearance customization had no significantly positive impact on the intrinsic motivation of crowd workers (with respect to interest-enjoyment and effort-importance). This can potentially be explained by the fact that workers are mainly motivated by monetary incentives in paid crowdsourcing marketplaces, rather than by the allure to stimulate their feelings of enjoyment and interest through task execution. Based on our findings in the paid microtask crowdsourcing setup, future work can explore the potential of avatar customization in voluntary crowdsourcing.

In terms of perceived workload as well as quality-related outcomes, the differences in Image Transcription tasks across conditions were not found to be statistically significant. Through open-ended comments at the end of the tasks, several workers reported that they found the Image Transcription tasks rather boring and repetitive, affecting their experience. Example comments reflecting these notions are shared below.

> *I would have done more, but captas (authors' note: CAPTCHAs) are really not my thing. Maybe have a choice between different types of tasks. Thanks.* (from a male worker in a pleasant mood)

> *I have done too many of these as security questions to make sure I am not a bot. This is how I am associating these. Makes it very uninteresting.* (from a male worker in a pleasant mood)

In contrast, workers found the Information Finding tasks to be interesting, since they provided workers with a chance to gain new knowledge and learn some interesting facts through the course of executing the tasks. Example comments reflecting these notions are shared below. It is interesting to note these perceptions despite the fact that Information Finding tasks required more effort and were time-consuming (Information Finding $64.6\pm59.8$ vs Image Transcription $35.9 \pm 24.7$, unit: seconds per microtask).

> *Some interesting trivia kind of like when you go to a random wikipedia page.* (from a male worker in a neutral mood)

> *Thanks for the task, it was cool searching and remembering some celebrities. The chatbot works very well!* (from a female worker in a pleasant mood)

Our findings suggest that the task type can influence workers' experience, and using worker avatars can be more effective in tasks that are more challenging and exploratory, especially when they involve elements or opportunities for learning [52].

In this study, the post-task survey was conducted on either the Chat interface or the Web interface depending on the experimental condition. However, a previous study has pointed out that a "casual" conversational style could influence participants' responses [113]. To avoid this potential confound, in our post-task surveys we used a "formal" conversational style (which is known to not influence worker responses). We emphasize the caveat that other complex factors may indeed affect the nature of responses when a conversational interface is employed.

In this study we did not investigate the impact of worker characterization selection independently (which implies another experimental condition — characterization selection without avatar appearance customization). This design choice was driven by our interest in understanding the impact of using avatars, and whether adding worker characterization can help increase workers' self-identification with their avatars. However, we will explore the impact of characterization independently in our imminent future work, as it would further our understanding of the interplay between worker avatars and characterizations.

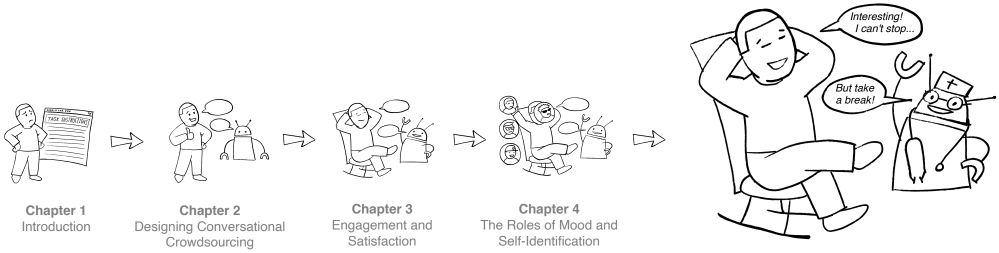
## 4.3 Chapter Summary

In this chapter, we studied how worker moods and self-identification could affect conversational crowdsourcing. In the first study (Section 4.1), we explored how worker moods could affect their output quality, engagement, and cognitive task load in conversational microtask crowdsourcing, and how the conversational style of the conversational agent could affect the performance of workers in different moods. We evaluated worker performance across different tasks by comparing quality related outcomes between different interfaces (and conversational styles). In the second study (Section 4.2), we investigated whether using worker avatars and enabling avatar customization through avatar appearance and characterization could improve worker experience in conventional web and novel conversational interfaces. We designed worker interfaces combining the avatar customization affordances. We carried out a between-subjects experimental study with 360 workers to analyze their perceived workload, intrinsic motivation, and the concomitant quality-related outcomes. We evaluated how worker avatar appearance and characterization could affect intrinsic motivation during task execution, and studied the traits of the workers who selected different characterizations.

We show the strong evidence that workers' subjective perceptions can affect conversational crowdsourcing. In terms of moods, we highlight that workers in a pleasant mood generally exhibit a better worker performance. The findings also suggest that a suitable conversational style can have a significant impact on engaging workers in different moods, which shows the opportunities for mood-aware task assignment to achieve better worker satisfaction and higher output quality in crowd work. In terms of self-identification, the results suggest using self-identification with worker avatars has positive effects in improving worker experience in conversational crowdsourcing with very small overheads. The findings provide important implications in improving requester-worker relationship and ensuring the sustainability of crowdsourcing marketplaces.

# Chapter 5

# Applications of Conversational Crowdsourcing


TASK INSTRUCTIONS
Interesting! I can't stop...
But take a break!
Chapter 1
Introduction
Chapter 2
Designing Conversational Crowdsourcing
Chapter 3
Engagement and Satisfaction
Chapter 4
The Roles of Mood and Self-Identification

Conversational crowdsourcing has been shown to be effective in improving worker engagement and satisfaction during microtasking. We also analyzed the roles of worker moods and self-identification in conversational crowdsourcing for executing microtasks. Since crowdsourcing has become a primary means for not only conventional microtasking but also carrying out online human-centered studies, the effectiveness of conversational crowdsourcing in general human-centered studies still remains unexplored. In this chapter, we introduce two applications of conversational crowdsourcing.

In Section 5.1, we first explored the possibility of applying conversational crowdsourcing in the field of information retrieval, to investigate the effect of conversation on human memorability due to the fact that information overload is a problem many of us can relate to nowadays. The deluge of user generated content on the Internet, and the easy accessibility to a vast amount of data compounds the problem of remembering and retaining information that is consumed. Previous works in online education have shown that conversational systems can improve learning effects. Although memorization is an important part of learning, the effect of conversation on human memorability remains unexplored. Furthermore, to make information consumed more memorable, strategies such as note-taking have been found to be effective by augmenting human memory under specific conditions. This is based on the rationale that humans tend to recall information better if they have produced the information themselves. We aim to address this knowledge gap through an experimental study using conversational crowdsourcing, by investigating human memorability in a classical information retrieval setup. We explore the impact of conversational interfaces and note-taking affordances on the memorability of information consumed by users. Our results showed that traditional web search and note-taking had positive effects on knowledge gain, while the search engine with a conversational interface had the potential to augment long-term memorability. This work highlights the benefits of using conversational interfaces and note-taking to aid human memorability.

In the second study (Section 5.2), we used conversational crowdsourcing to understand the worker health status across different crowdsourcing platforms, since the health and wellbeing of crowd workers has become an important concern with the growing landscape of online work in general and the rise of paid microtask crowdsourcing in particular. A substantial amount of prior work has explored challenges pertaining to improving the effectiveness and efficiency of microtask crowdsourcing from the standpoint of quality. Only a few works however, have attempted to address the pertinent concerns of crowd workers' health and wellbeing. In contrast to traditional work settings where employee health is safeguarded by contractual laws and obligations, the unregulated dynamics of microtask crowdsourcing marketplaces expose crowd workers to a multitude of potential health-related risks. Though recent work has highlighted issues pertaining to the unfair treatment of crowd workers and the abysmal pay for piecework, little is currently understood about worker health and well-being on crowdsourcing platforms. In this study, we used conversational crowdsourcing to deploy a 60-item survey on two popular crowdsourcing platforms – Amazon Mechanical Turk and Prolific – to better understand workers' health-related background, physical health, mental health, and their needs. We found that workers across these platforms reported similar health-related issues, but also exhibited certain differences. Based on our findings, we argue that crowdsourcing platforms, task requesters, and academic researchers need to take the collective responsibility of creating better work environments and ensuring worker wellbeing. We argue that improving worker health on crowdsourcing platforms is the crucial need of the hour to ensure a sustainable future for crowd work.

The findings from this chapter have important implications in helping users better retain information acquired from computer systems, and on task and workflow design that are centered around worker health on crowdsourcing platforms. The experiments carried out in this chapter showed that conversational crowdsourcing could be an effective tool for performing human-centered studies.

The content of this chapter is based on the following papers:

Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon. Towards Memorable Information Retrieval. Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval, pp. 69-76, 2020. (Section 5.1 is based on this paper)

Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon. Understanding Worker Health on Crowdsourcing Platforms. A new paper submitted to CSCW 2021. (Section 5.2 is based on this paper)

## 5.1 Towards Memorable Information Retrieval

Information overload is a byproduct of the rapid development of information technology and the plethora of user generated content. By issuing a simple search query, an Internet user can access billions of relevant items from a search engine within seconds. The data deluge and a constant exposure to new information leads to the problem of remembering and retaining information during informational search sessions. Most popular search engines today are optimized to serve relevance related needs with respect to user queries. We believe that an unexplored opportunity lies in how information can be retrieved and presented to users, with an aim to improve the memorability of information consumed.

To improve human memorability, researchers in the field of experimental psychology have studied the "generation effect" [199]. By comparing memory for words, experiments revealed that humans could better recall information if they produced it themselves rather than if they received it. Based on the generation effect, prior studies have shown that note-taking, a simple way to re-produce received information, can improve human memorability, particularly for text-based learning and comprehension [24, 200]. However, the effects of note-taking in a classic information retrieval setup remain unexplored.

Prior studies in online learning have revealed that conversational systems can significantly improve learning outcomes [90, 129, 203]. As the goal of learning is to develop a deep understanding of some information, memorization is an important element [111, 12]. Although conversation can produce unique context linked with information, the effect of conversational systems on human memorability needs further exploration. In Chapter 3, we have investigated the role of text-based conversational interfaces in online information finding tasks. We demonstrated that a conversational interface could better engage online users. However, the question of whether improved user engagement through conversational interfaces leads to better memorability of information remains unanswered. We aim to fill this knowledge gap by proposing novel approaches to improve human memorability during information retrieval. We specifically focus on information retrieval activities carried out through the Web search using desktop browsers. Through rigorous experiments, we seek to address the following research questions.

**RQ5.1:** How can human memorability of information consumed in informational web search sessions be improved?

**RQ5.2:** How does the use of text-based conversational interfaces and note-taking affect the search behavior of users?

Inspired by prior work in psychology and HCI, we propose novel search interfaces which (a) provide the affordance of note-taking to users, and (b) provide a conversational interface. We propose methods to quantify knowledge gain and long-term memorability of information consumed, and investigate the impact of the proposed search interfaces on the memorability of information consumed. We conducted an online user study in a classical information retrieval setup. Results revealed that traditional Web interfaces with a note-taking affordance could benefit knowledge gain (up to 25% higher than other interfaces), while conversational interfaces had the potential to augment long-term memorability (7.5% lower long-term information loss). Furthermore, we found that users leveraging conversational interfaces input more queries but opened links less frequently compared to users leveraging the traditional Web interfaces. In addition, the users of conversational interfaces tended to type notes themselves, while the Web users input significantly longer notes by copying content directly from the search engine result pages. Our findings suggest that both note-taking and conversational interfaces are promising tools for augmenting human memorability in information retrieval.

## Related Work: Human Memorability and Information Systems

### Augmenting Human Memory

Different theories for augmenting human memory have been studied in the field of psychology. The memory consolidation theory proposed by Müller and Pilzecker explained the processes to make information memory [155, 147]. The Atkinson-Shiffrin memory model shows that the long-term memory can be consolidated by repeatedly rehearsing short-term memory [4]. To study how the 'remembering information' relates to one's self, previous work has revealed that the memory could be enhanced if it relates to one's self-concept or an episode from one's life [17]. A prior study in experimental psychology has shown evidence of the existence of the "generation effect" [199]. Authors conducted experiments at the word-level to show that people could remember information better if the information was produced by themselves. A simple and direct application of the generation effect is the use of note-taking. Previous studies have shown that note-taking can improve human memorability in different scenarios [56, 149, 24, 200]. Intons-Peterson et al. examined the use of internal and external memory aids in experiments with 489 undergraduates. It was found that at least one external aid, i.e. taking notes, can effectively facilitate remembering [103]. Based on the findings of prior works, in this study we investigate how an external aid such as note-taking can affect the long-term memorability of users in informational search.

### Aiding Memorability in Information Systems

Augmenting human memory has also been studied from an information systems standpoint. Many previous studies have used context as a key aspect to improve human memorability [179, 47]. The 'Remembrance Agent' is an automatic system which uses the role of context in memory to augment human memory, by listing documents related to the user's current context [179]. Blanc-Brude et al. have performed experiments to find the attributes (e.g. file name, time, title, location, size, etc.) that help memorability for a document search tool [16]. Previous works have also shown that many strategies, such as time-aware contextualization [28, 219], and optimizing recollection by generating analogies [180], have a positive effect on human memorability. Furthermore, a recent study built an application named 'ReflectiveDiary', to investigate how self-generated daily summaries can improve memorability [185]. Predictive methods have also been proposed to consolidate human memory in the workplace environment [8]. Since memorization is an essential element of the learning process [111, 12], we also examined relevant literature in online learning. Across multiple studies, conversational systems were found to be useful in facilitating learning effects [90, 129, 203]. In Chapter 2, we showed that conversational interfaces could effectively improving user engagement in information retrieval tasks. These previous works with regard to aiding memorability or improving learning effects in information systems are not directly applicable in the current information retrieval ecosystems. Inspired by these prior works, we propose novel search interfaces and design experiments to study human memorability in information retrieval.

## Method

The goal of this study is to investigate whether note-taking and conversational interfaces can affect human memorability in informational web search sessions. To this end, we measure long-term memorability of information consumed by users.

### Study Design

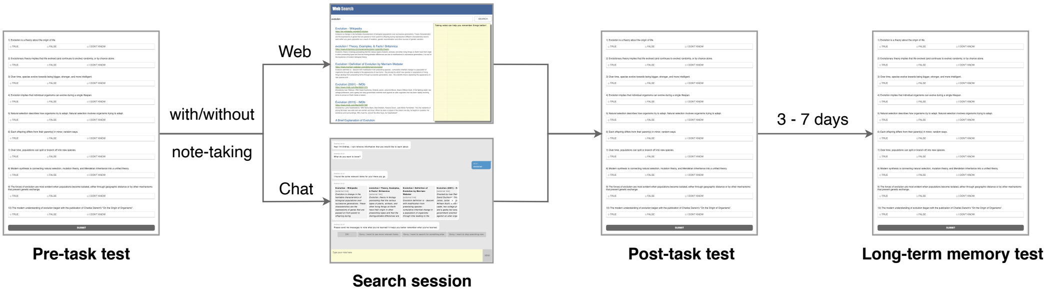


Figure 5.1: Workflow of our study. The *pre-task test*, the *search session* and the *post-task test* pertain to a single Human Intelligence Task (HIT) published on Amazon MTurk. The *long-term memory test* is deployed separately in a follow-up HIT.

Table 5.1: Topics and corresponding information needs (topics are re-used from [76]).

| ***Topic*** | **Information Need** |
|---|---|
| *Altitude Sickness* | The users are required to acquire knowledge about the symptoms, causes and prevention of altitude sickness. |
| *American Revolutionary War* | The users are required to acquire knowledge about the 'American Revolutionary War'. |
| *Carpenter Bees* | The users are required to acquire knowledge about the biological species 'carpenter bees'. How do they look? How do they live? |
| *Evolution* | The users are required to acquire knowledge about the theory of evolution. |
| *NASA Interplanetary Missions* | The users are required to acquire knowledge about the past, present, and possible future of interplanetary missions that are planned by the NASA. |
| *Orcas Island* | The users are required to acquire knowledge about the Orcas Island. |
| *Sangre de Cristo Mountains* | The users are required to acquire knowledge about 'Sangre de Cristo' mountain range. |
| *Sun Tzu* | The users are required to acquire knowledge about the Chinese author Sun Tzu - about his life, his writings, and his influence to the present day. |
| *Tornado* | The users are required to acquire knowledge about the weather phenomenon that is called 'tornado'. |
| *USS Cole Bombing* | The users are required to acquire knowledge about the 2000 terrorist attack that came to be known as the 'USS Cole bombing'. |

The taxonomy of human memory, which is rather complicated and detailed, has been developed for over a hundred years. Human memory can be classified into two big categories; short-term and long-term memory. Short-term memory only persists for seconds or minutes [7, 4, 80], while long-term memory can last for much longer [4]. In this study, we focus on improving the long-term memorability of information consumed by users in web search sessions. According to Ebbinghaus' curve and recent replication works [157]: the forgetting curve goes down slowly after 24 hours (people forget more than 60% within 24 hours, 70% within 2 days, and 80% within 30 days). It was found that fluctuations might appear at the 24-hour point. However, after 2 days, the forgetting curve becomes stable. Therefore, we choose 3-7 days as the time interval to measure user long-term memorability in this study.

The basic idea of measuring memorability in web information retrieval is to quantify how much information a user can remember at the end of an informational search session. Therefore, as shown in Figure 5.1, we first assign a topic and an information need to users, and ask the users to finish a "knowledge calibration" test (*pre-task test*) with 10 questions related to the topic. We use 10 topics and the corresponding questions from a previous work about analyzing knowledge gain in informational search [76], as listed in Table 5.1. Topics are randomly assigned to users. Through the pre-task test users can better understand different facets of the information need, and we can calibrate the background knowledge of users.

Next, users are directed to the search session, where they must spend at least 7 minutes searching about their assigned information need. As we can see from Figure 5.1, users are assigned any 1 of 4 different user interfaces. Half of the users use a Web interface to perform their search sessions, while the rest are assigned a conversational interface. Both Web and conversational interfaces have two conditions, i.e. with note-taking function enabled or disabled. In the Web interfaces, users leverage a Web search page that is similar to typical search engines. In the conversational interfaces, users are guided by a conversational agent through their session.

After the search session, users need to finish a *post-task test.* The questions shown in the *post-task test* are identical to the questions in the *pre-task test*, allowing us to measure user knowledge gain. To incentivize active search behavior during the search session, users were informed that an extra reward will be given depending on the number of correct answers in the *post-task test.* To elicit honest and genuine responses, users were also told that their accuracy in the *pre-task test* would not affect the reward.

Three days after the search session, we notify all the users who participated in our study and give them an opportunity to answer our *long-term memory test* within the next 4 days in return for an additional reward of 1 USD. The questions in the *long-term memory test* are identical to the *pre-task test.* By comparing the results of the *post-task test* to the *long-term memory test*, we can measure how much information users have retained or forgotten over this long-term period.

### Measuring Memorability

**Measuring knowledge gain.** Similar to prior work in *search as learning* [76, 240], we measure the knowledge gain of users as the normalized difference in performance of users between the post-task and pre-task knowledge tests.

We use $A_t$ ($t \in \{pre, post, long\}$) to denote the set of answers of the test $t$, and use $A_t^i \in A_t$ ($1 \leq i \leq 10$) to represent if the $i^{th}$ question of the test $t$ is correctly answered ($A_t^i = 1$) or not ($A_t^i = 0$) by the user. If a user chooses "I DON'T KNOW", we consider it as incorrect answer. For instance, if the $5^{th}$ question of the *pre-task test* is correctly answered by the user, then we assign $A_{pre}^5 = 1$; if the answer of the $7^{th}$ question of the *post-task test* provided by the user is incorrect, we assign $A_{post}^7 = 0$. Thus, the normalized knowledge gain can be calculated by using the following equation (where the max/min(*topic score*) means the maximum or minimum score among all the tests sharing the same topic, and the score of a test $t$ can be calculated by $\sum_{i=1}^{10} A_t^i$).

$$knowledge\ gain = \frac{\sum_{i=1}^{10} A_{post}^i - \sum_{i=1}^{10} A_{pre}^i}{\max(topic\ score) - \min(topic\ score)} \tag{5.1}$$

**Measuring long-term memorability.** Similarly, we can also use *information gain* to measure the long-term user memorability, which can be calculated by the following equation.

$$information\ gain = \frac{\sum_{i=1}^{10} A_{long}^i - \sum_{i=1}^{10} A_{pre}^i}{\max(topic\ score) - \min(topic\ score)} \tag{5.2}$$

Long-term memorability can also be measured using *information loss*. The *information loss* after the post-task test can be quantified by the number of questions which are correctly answered in the *post-task test* but incorrectly answered in the *long-term memory test*. Thus, it can be calculated by the following equation.

$$information\ loss = \frac{\sum_{i=1}^{10} A_{post}^i - \sum_{i=1}^{10} A_{long}^i \cdot A_{post}^i}{\max(topic\ score) - \min(topic\ score)} \tag{5.3}$$

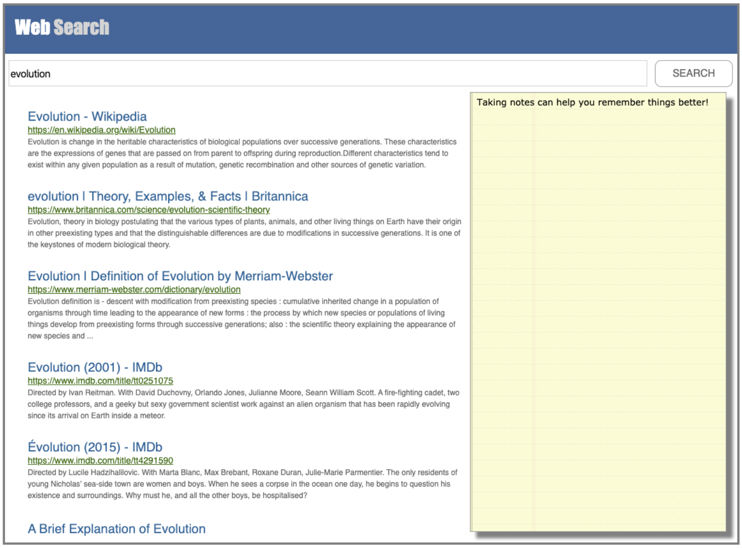
### User Interfaces

Addressing **RQ5.1**, we designed Web and conversational interfaces to support informational search sessions, with an optional note-taking functionality. Both the Web and conversational interfaces use the Bing Search API [32] for sending search query requests and receiving search results (relevant web pages).

**The Web interface** is designed according to the typical user interface of popular search engines, as shown in Figure 5.2 (a). The Web interface consists of two main components — a textarea for entering search queries, and a rectangular frame for displaying search results. During the search session, users need to type search queries in the textarea at the top of the page. Users can either click the "SEARCH" button or press the "Enter" key on the keyboard to issue the search query asking for 10 relevant items (Web pages), and then the sever will respond with a list of search results. The search results include 10 items with their titles, links and snippets, which are shown under the text area, occupying the most part of the Web interface. Since each query fired only requests for 10 relevant items, the Web interface only shows 10 search results at a time. Each item is clickable. To prevent users from jumping to other pages or applications, once the user clicks an item, an embedded browser will pop up to show the content of the corresponding item (Web page). To retrieve more items, users can click the "NEXT PAGE" button to send a query asking for the next 10 relevant items, or click the "PREVIOUS PAGE" button to go back.

Furthermore, as shown in Figure 5.2 (b), to enable the function of note-taking, a notepad is embedded on the right side of the Web interface. The notepad can be enabled or disabled depending on the experimental condition. On the notepad, we leave a sentence "tasking notes can help you remember things better" to encourage users to take notes during the search session. All the on-page activities including querying, browsing (clicking) items, and note-taking are automatically logged for user behavior analysis.

[32] `https://www.customsearch.ai/`



Web Search

evolution

SEARCH

Evolution - Wikipedia
https://en.wikipedia.org/wiki/Evolution
Evolution is change in the heritable characteristics of biological populations over successive generations. These characteristics are the expressions of genes that are passed on from parent to offspring during reproduction.Different characteristics tend to exist within any given population as a result of mutation, genetic recombination and other sources of genetic variation.

evolution | Theory, Examples, & Facts | Britannica
https://www.britannica.com/science/evolution-scientific-theory
Evolution, theory in biology postulating that the various types of plants, animals, and other living things on Earth have their origin in other preexisting types and that the distinguishable differences are due to modifications in successive generations. It is one of the keystones of modern biological theory.

Evolution | Definition of Evolution by Merriam-Webster
https://www.merriam-webster.com/dictionary/evolution
Evolution definition is - descent with modification from preexisting species : cumulative inherited change in a population of organisms through time leading to the appearance of new forms : the process by which new species or populations of living things develop from preexisting forms through successive generations; also : the scientific theory explaining the appearance of new species and ...

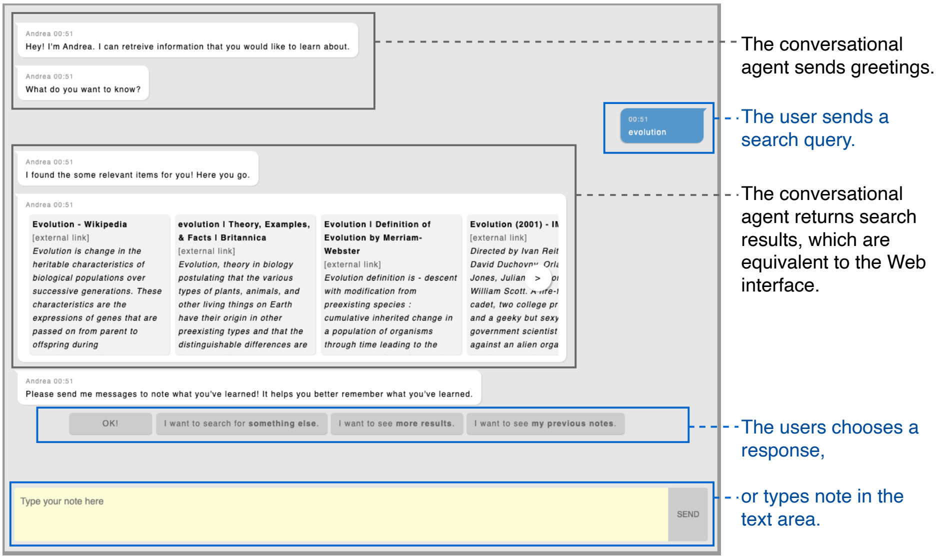
Evolution (2001) - IMDb
https://www.imdb.com/title/tt0251075
Directed by Ivan Reitman. With David Duchovny, Orlando Jones, Julianne Moore, Seann William Scott. A fire-fighting cadet, two college professors, and a geeky but sexy government scientist work against an alien organism that has been rapidly evolving since its arrival on Earth inside a meteor.

Évolution (2015) - IMDb
https://www.imdb.com/title/tt4291590
Directed by Lucile Hadzihalilovic. With Marta Blanc, Max Brebant, Roxane Duran, Julie-Marie Parmentier. The only residents of young Nicholas' sea-side town are women and boys. When he sees a corpse in the ocean one day, he begins to question his existence and surroundings. Why must he, and all the other boys, be hospitalised?

A Brief Explanation of Evolution

Taking notes can help you remember things better!

Figure 5.2: Web search interfaces with the note-tasking function enabled. The yellow notepad becomes invisible if the note-taking function is disabled.



Andrea 00:51
Hey! I'm Andrea. I can retreive information that you would like to learn about.

Andrea 00:51
What do you want to know?

00:51
evolution

Andrea 00:51
I found the some relevant items for you! Here you go.

Andrea 00:51

**Evolution - Wikipedia**
[external link]
*Evolution is change in the heritable characteristics of biological populations over successive generations. These characteristics are the expressions of genes that are passed on from parent to offspring during*

**evolution | Theory, Examples, & Facts | Britannica**
[external link]
*Evolution, theory in biology postulating that the various types of plants, animals, and other living things on Earth have their origin in other preexisting types and that the distinguishable differences are*

**Evolution | Definition of Evolution by Merriam-Webster**
[external link]
*Evolution definition is - descent with modification from preexisting species : cumulative inherited change in a population of organisms through time leading to the*

**Evolution (2001) - IM**
[external link]
*Directed by Ivan Reit David Duchovny Orla Jones, Julian William Scott. cadet, two college pr and a geeky but sexy government scientist against an alien orga*

Andrea 00:51
Please send me messages to note what you've learned! It helps you better remember what you've learned.

>
OK!
I want to search for **something else**.
I want to see **more results**.
I want to see **my previous notes**.
Type your note here
SEND

The conversational agent sends greetings.

The user sends a search query.

The conversational agent returns search results, which are equivalent to the Web interface.

The users chooses a response,

or types note in the text area.

Figure 5.3: Conversational search interface.

**The conversational interface** uses the same search engine as the Web interface and is implemented based on TickTalkTurk. However, the search workflow is guided by a text-based conversational agent, as shown in Figure 5.3. The logic of the conversational interface for web search is designed as follows:

*1) Greetings.* The conversational agent opens the conversation with the user and then asks the user to provide a search query. The conversational agent sends the greetings to initiate the search session.

> *- Hey! I'm Andrea. I can retrieve information that you would like to learn about.*
> *- What do you want to know?*

Note that we assign a gender-neutral name ("Andrea") to the conversational agent, to avoid potential biases. Andrea is a name commonly used for both males and females around the world.

*2) Search.* After the user provides the agent with a search query, the conversational agent uses Bing Search API to retrieve results. To make the conversational interface comparable to the Web interface, the agent also shows 10 relevant items at a time. However, on the conversational interface, all the content is presented within chat bubbles to replicate typical conversational interfaces. As we can see from Figure 5.3, the relevant items are listed horizontally in a chat bubble, where the user can scroll horizontally to view them. Also, each item in the chat bubble is clickable and linked to the embedded Web browser.

*3) Response selection.* The conversational agent provides the user with four options after the search results have been displayed. The four options correspond to taking notes, showing more results, entering a new query, and showing previous notes, respectively. However, if the note-taking function is disabled, the agent only presents two options — showing more results and entering a new search query. If a user chooses to **take notes**, the message that the user sends to the agent will be recorded and integrated with previous notes (if any) from the user in the search session. If the user chooses **show more results**, the next 10 relevant items will be displayed to the user with a new chat bubble. The functionality is equivalent to that of the Web interface. The conversational interface does not provide an option to show previous items, since users can easily find previous items by viewing the conversation history. If the user chooses to **input a new query**, the agent goes back to *step 2 Search* to re-start the search process. Finally, all the previous notes can be shown in a chat bubble if the user chooses to see the notes by using the **show previous notes** option.

## Experiments

### Experimental Conditions

In this study, we use two user interfaces (Web and Conversational) with a note-taking function either enabled or disabled to address our research questions. This results in four experimental conditions.

**Chat w/ note:** the conversational interface with note-taking. In this experimental condition, users are redirected to a conversational interface, where the searching process is guided by a conversational agent — Andrea. In addition, the note-taking function is enabled, meaning users can take notes by sending messages to Andrea.

**Chat w/o note:** the conversational interface without note-taking. In this experimental condition, users are redirected to an ordinary conversational interface, where the searching process is also guided by Andrea, but the note-taking function is disabled.

**Web w/ note:** the Web interface with note-taking. In this experimental condition, users are redirected to a custom Web search interface to complete the search session. A notepad is visible on the right side of the Web interface where users can type their notes.

**Web w/o note:** the Web interface without note-taking. In this experimental condition, users are also redirected to a custom Web search interface to complete the search session. However, the notepad is hidden and disabled. This experimental condition represents the most typical search engines nowadays.

Participants in our experimental study were recruited from Amazon Mechanical Turk (MTurk). The code along with all the data are made available to the community to facilitate further research[33]. We published online tasks with the aforementioned four experimental conditions on MTurk. The Human Intelligence Task (HIT) published on MTurk only contained the *pre-task test*, search session and the *post-task test*. The *long-term memory test* was not included in the HIT batches. We used the notification function provided by MTurk, to send the link of the *long-term memory tests* to workers after three days. The Web page of the *long-term memory test* was set up on our own server. We recruited 35 online crowd workers per condition from MTurk, as the users of our search systems. Each worker was assigned a random topic from Table 5.1. The experiment was approved by the ethics committee of TU Delft, and we did not collect and store any identifiable data of human subjects.

### Quality Control

The minimum time for each search session was set to 7 minutes (users were not allowed to proceed to the next stage before 7 mins). Apart from incentivizing genuine search behavior through attached rewards for performance in the *post-task test*, we took additional measures to ensure reliable behavior. The timer stops if a worker temporally leaves the page (for instance, switching to other tabs or programs). Furthermore, we use an embedded browser to enable workers to open and browse the search results on our own task page, instead of opening a new tab. Considering the effects of learning bias, we add an extra Javascript code to record the unique MTurk Worker ID on our server, to prevent a worker from executing our HIT multiple times. We restricted participation by using the default qualification type, "Overall HIT approval rate is greater than 95%" provided by MTurk to further ensure high worker quality. In addition, we manually inspected users' answers to exclude any potentially unreliable users. We exclude users if they:

1. Enter no queries during the search session;

2. Always select the same option — either 'YES' or 'NO' in *pre-/post-task test* or *long-term memory test*.

Due to the criteria we defined, 8 workers were manually excluded in our experiments.

### Worker Reward

Upon the task completion, we immediately reward each worker with 2 USD. After three days (72 hours), we bonus workers according to the number of correct answers given in the *post-task test* (0.01 USD per correct answer). In the notification message corresponding to the bonus, we requested workers to participate in our *long-term memory test* by providing a link to the test page. The Web page of the *long-term memory test* is set up on our own server instead of MTurk. We incentivized workers to complete the *long-term memory test*

[33] `https://sites.google.com/view/memorableir`

with an additional reward of 1 USD on completion. For the next three days (i.e., until 7 days after their search session), we sent a notification every 24 hours to those workers who did not finish *the long-term memory test* yet.

### Evaluation Metrics

We measure the user knowledge gain, long-term memorability, search time, and user behavior including number of queries, browsing frequency and the length of notes that users take (where applicable) while completing the HITs.

**Knowledge Gain and Long-term Memorability.** User knowledge gain is calculated using Equation 5.1. The long-term memorability is measured using (i) information gain, calculated using Equation 5.2, and (ii) information loss, calculated using Equation 5.3.

**Search time.** We recorded how long each user spends on the search session, which is the length of the time period starting from when the user submits the answers of the *pre-task test*, until the worker clicks the “NEXT” button to proceed to the *post-task test.* The “NEXT” button becomes visible only after 7 minutes, enforcing a minimum search time of 420 seconds.

**Search Behavior.** We also analyze user behavior during the search sessions to better understand how user behavior relates to the memorability of information consumed. To this end, we focus on:

*1) Number of queries.* It represents how many queries a user sends to search engine through either Web or conversational interfaces;

*2) Browsing frequency.* This is the frequency of a user opening a link and using the embedded Web page browser to view the content of the search results.

*3) Length of notes.* It represents the number of characters written in the notes provided by the user.

## Results

After excluding unreliable workers, the four experimental conditions — **Chat w/ note**, **Chat w/o note**, **Web w/ note**, and **Web w/o note**, we are left with 32, 34, 33, and 33 unique valid users respectively. Furthermore, the four conditions had 14, 11, 15 and 16 users who returned for the *long-term memory test* respectively.

### Memorability Analysis

**Knowledge gain.** *The Web interface with note-taking can significantly improve the knowledge gain in comparison to the conversational interface conditions, while the conversational interface without note-taking shows no positive impact on the knowledge gain of users.*

Figure 5.4 presents the knowledge gain of users across the four interface conditions. The average knowledge gain of users corresponding to the conversational interfaces is 4.4%, while that of the Web interfaces is 21.5%. Particularly, the knowledge gain of the Web interface with note-taking function enabled (Web w/ note) is 25% higher than the conversational interface with note-taking function disabled. Since the distributions of knowledge gain follow normal distributions (verified by the Shapiro-Wilk tests for normality), we use independent t-tests ($\alpha = 0.05$) to find the significant differences between user interfaces. We found three pairs having a $p$-value less than 0.05 (Chat w/ note vs Web w/ note $p$=0.030, Chat w/o

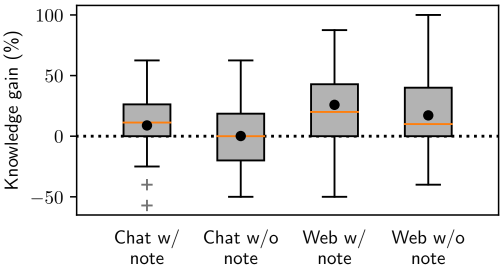


Figure 5.4: Knowledge gain of users across the four interfaces.

note vs Web w/ note $p$=9.7e-4, and Chat w/o note vs Web w/o note $p$=0.031). After Holm-Bonferroni correction, the knowledge gain of Web w/ note is still significantly higher than Chat w/o note. Results suggest that note-taking is a useful tool for improving knowledge gain, aligned with findings from previous studies. However, the conversational interface revealed no specific advantage over the traditional web interface in facilitating knowledge gain.

**Long-term Memorability.** *Results revealed no significant difference across interface conditions with regard to long-term information gain (computed using information gain). However, conversational interfaces exhibit the potential to reduce long-term information loss.*

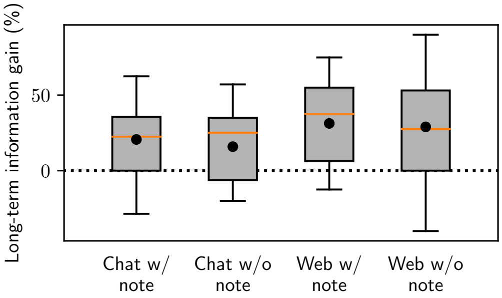


Figure 5.5: Long-term memorability (using information gain) across the four interfaces.

As shown in Equation 5.2, the average long-term information gain of users across all user interfaces is actually higher than the average knowledge gain observed. This is due to the subset of users who returned to complete the *long-term memory test* (these users had relatively higher knowledge gain scores). We also found that the long-term information gain is significantly correlated to knowledge gain according to Pearson correlation coefficient testing ($p < 0.05$ except Chat w/o note). The distributions of long-term information gain also follow normal distributions (verified by the Shapiro-Wilk test for normality). However, we found no significant difference between long-term information gain across the four interface conditions by independent t-tests. This suggests that the Web interface and note-taking show no positive effect on long-term memorability, although they can effectively improve knowledge gain.

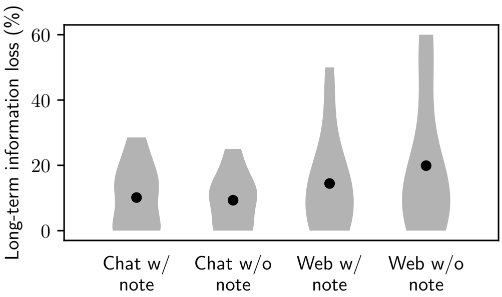


Figure 5.6: A violinplot of long-term memorability (information loss) across the four interfaces.

The long-term information loss was calculated to further analyze long-term memorability. The distributions of information loss across four interfaces are shown in Equation 5.3. We found the average information loss of users corresponding to conversational interfaces is (9.8%), which is 7.5% lower than that of the Web interfaces (17.3%) with a small $p$-value ($p = 0.06$, independent t-tests). Furthermore, the maximum information loss among the 25 users who use conversational interfaces is 28%, while that of the 31 users using Web interfaces is 60%. These results indicate that the conversational interface has the potential to improve user long-term memorability.

### Search Time Analysis

**Search Time (in seconds)**. *We found no significant difference in the average search time of users across the four interface conditions.*

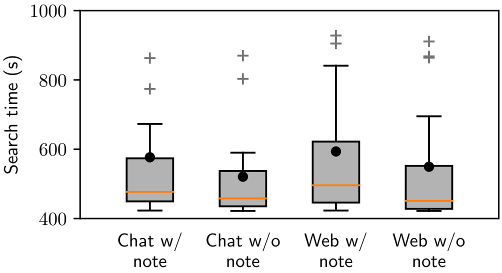


Figure 5.7: Search time (in seconds) across the four interface conditions.

We measured the time that the user spends on the search session for each experimental condition. The average search time across all user interfaces is 559 seconds. As the distributions of search time do not follow normal distribution (according to Shapiro-Wilk tests), we used Mann-Whitney U tests ($\alpha = 0.05$) to compare the search time across four user interfaces. Although the average search time of the user interfaces with note-taking, for both conversational and Web interface, is slightly higher than the user interfaces without note-taking, this was not found to be statistically significant.

### Worker Behavior Analysis

The worker behavior during the search session is analyzed using three measurements, i.e. the number of queries, the browsing frequency, and the length of notes.

Table 5.2: Mean and standard deviation ($\mu \pm \sigma$) of the number of queries, the browsing frequency, and the length of notes across the four user interface conditions.

| *Interfaces* | **Number of queries** | **Browsing frequency** | **Length of notes** |
|---|---|---|---|
| *Chat w/ note* | $9.56 \pm 5.23$ | $0.47 \pm 1.09$ | $348.68 \pm 457.15$ |
| *Chat w/o note* | $9.71 \pm 8.66$ | $0.44 \pm 1.01$ | / |
| *Web w/ note* | $3.76 \pm 3.16$ | $1.82 \pm 2.43$ | $1004.58 \pm 1431.63$ |
| *Web w/o note* | $4.64 \pm 5.66$ | $2.09 \pm 1.96$ | / |

**Number of Queries**. *The users corresponding to conversational interfaces tend to send more queries on average (ask more questions to the conversational agents), while the users corresponding to the web interfaces input significantly fewer queries.*

In terms of the number of queries, we found that users using conversational interfaces generally send more queries than the users who use Web interfaces (2.3 times more queries). We applied Mann-Whitney U tests ($\alpha = 0.05$) and Holm-Bonferroni correction to discover significant differences across conditions with respect to number of queries fired. Results of significant testing revealed that note-taking had no impact on the number of queries. However, the conversational interfaces significantly increase the number of queries entered by users, compared to the traditional Web interfaces (Chat w/ note vs Web w/ note $p =$ 3.5e-6; Chat w/ note vs Web w/o note $p =$ 2.6e-5; Chat w/o note vs Web w/ note $p =$ 1.6e-5; Chat w/o note vs Web w/o note $p =$ 9.1e-5). Moreover, a manual investigation of the search histories show that users using conversational interfaces tend to use questions as queries. This suggests that the users using a conversational interface tend to retrieve information by frequently posing questions to the agent as expected.

**Browsing frequency**. *Users in the conversational interface conditions tend to retrieve information by viewing snippets rather than by frequently opening links.*

The browsing frequency represents the frequency with which a user opens the links of search results. We found that note-taking has no significant impact on the browsing frequency of users according to Mann-Whitney U tests, while users using Web interfaces depict a significantly higher frequency of browsing search results (Chat w/ note vs Web w/ note $p =$ 0.0013; Chat w/ note vs Web w/o note $p =$ 4.9e-05; Chat w/o note vs Web w/ note $p =$ 9.0e-04; Chat w/o note vs Web w/o note $p =$ 3.8e-05). Our results suggest that the users using Web interfaces open the links more frequently, while the users of conversational interfaces tend to obtain information from snippets. This behavior of users in conversational interfaces can potentially be explained by their reluctance to break the coherence of conversation by opening links.

**Length of Notes**. *The users corresponding to web interfaces input significantly longer notes by copy-pasting content directly from the source, while the users in the conversational interface conditions type shorter notes by themselves.*

As for the length of notes, the users in the Web interface conditions input significantly longer notes compared to the users in the conversational interface conditions ($p$=0.022, Mann-Whitney U tests). A manual inspection reveals that users of web interfaces prefer copying content from the search results and pasting it to the notepad, while the users of conversational interfaces tended to type information themselves. Prior work has revealed that generating information by onself (notes), can aid long-term memorability. The fact that users in the conversational interface conditions indulged in generating notes themselves is promising and should be explored in future work.

**Worker Behavior and Long-term Memorability**. We investigated the linear relationship between users' search behavior and the memorability of the information consumed across the four interface conditions.

We performed Pearson correlation coefficient testing ($\alpha = 0.05$) to find the potential correlation between long-term memorability and all the worker behavior measurements. Although no statistical significance was found after Holm-Bonferroni correction, here we report the pairs whose $p$-value is less than 0.2. We found that the information loss has negative correlations with the number of queries and the length of notes, for users using a conversational interface with note-taking ($R = -0.46, p = 0.10$ and $R = -0.43, p = 0.13$ respectively). This indicates that the greater the number of queries or the longer notes that a user inputs, the less information the user tends to forget. As for users using a Web interface with note-taking, we found the information loss has positive correlations with the number of queries and the the browsing frequency ($R = 0.48, p = 0.07$ and $R = 0.58, p = 0.02$ respectively), indicating that a higher frequency of querying and browsing can potentially lead to information loss on a Web interface.

### Design Implications

Our findings in this study revealed that users employing conversational interfaces in informational search sessions exhibited a different search behavior compared to traditional web search: they relied primarily on text-based conversation, resulting in a significantly higher frequency of issuing queries but significantly lower frequency of opening SERP (search engine results page) links. This can potentially explain the relatively lower knowledge gain corresponding to users in the conversational interface conditions, since these users appear to consume information by means of viewing titles and snippets rather than opening links and exploring SERPs in detail. In contrast, our results indicated that note-taking in the traditional web interface could significantly increase user knowledge gain. These findings suggest that both note-taking and conversational interfaces can be promising tools towards achieving memorable information retrieval.

Furthermore, we found that users employing conversational interfaces had the potential to better retain information consumed (conversational interfaces were found to reduce long-term information loss). This is possibly due to the fact that conversational interfaces can generate unique context connected to the information during the search session. Our inspection of users' notes also corroborate that users using conversational interfaces tend to generate the information by themselves rather than copying content from sources (Web users' preference). In summary, this study demonstrated the feasibility of using conversational crowdsourcing in the field of information retrieval for carrying out user studies related to the use of conversational interfaces or agents.

### Limitations and Future Work

In this study, we found that using note-taking and conversational interfaces could enhance human long-term memory, and the users tended to exhibit different subjective perceptions. Therefore, to what extent the note-taking with different perceptions can improve (or probably reduce) information retrieval performance needs further exploration.

We found that only around half of the users returned for our *long-term memory test*, which is typical of such experiments. Our results showed that the users with a relatively higher *post-task test* scores were more willing to return and participate in our *long-term memory test*. It should be noted that this participation bias presents a threat to the representativeness of our findings. In our imminent future research on memorable information retrieval, we will explore whether a higher user engagement relates to a better user memorability of information consumed.

## 5.2 Understanding Worker Health

Microtask crowdsourcing marketplaces have thrived over the last decade due to the growing demand for accessible and cost-effective human input. Despite the central role that microtask crowdsourcing plays in research and industry, relatively little has been done to understand, improve, and safeguard the health and wellbeing of crowd workers, who form the very backbone of this paradigm. Commendable efforts have been devoted to create an awareness about the invisible labor that prevails on microtask crowdsourcing platforms [83]. Prior works have proposed tools to help crowd workers address issues related to power asymmetry and worker invisibility on platforms [104, 105, 197]. Researchers have highlighted the impact of work rejection and the importance of facilitating trust between workers and requesters [148, 190]. Others have built platforms to facilitate collective action [189], and proposed methods to help crowd workers share their risks alongside rewards on crowdsourcing platforms [59]. Recent work has also proposed methods to automatically ensure fair pay for crowd workers [229], and reduce the negative impact of exposure to harmful content [44]. However, little is currently understood about the status quo of overall workers' health on crowdsourcing platforms.

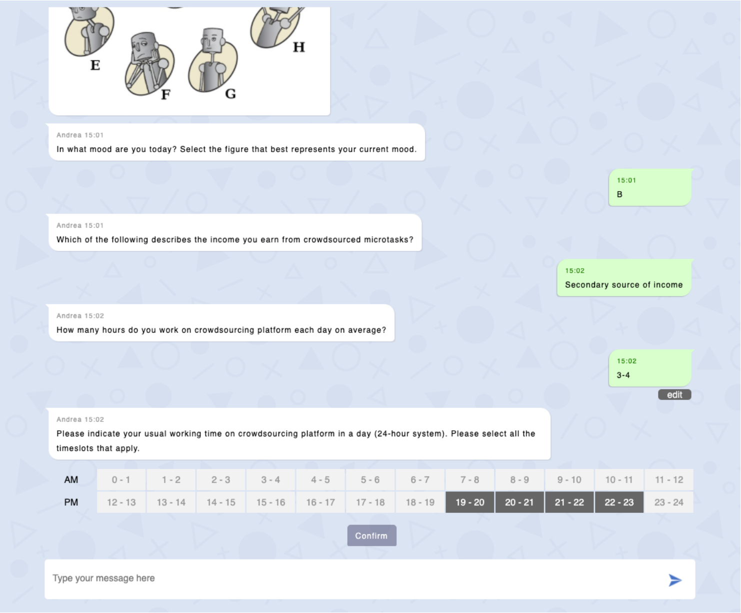
In contrast to traditional modes of work and employment where employee health is safeguarded to varying degrees through contractual laws and obligations, on-demand crowd work on crowdsourcing platforms currently lack such provisions for workers. Characterizing such '*ghost work*' recently, Gray and Suri [83] describe crowd work as a revolving door of temporary tasks, devoid of long-term prospects and guarantees. The unregulated dynamics of microtask crowdsourcing marketplaces with regard harmful content, power asymmetry and invisible labor expose crowd workers to a multitude of potential health-related risks [83, 190, 204]. We aim to better understand worker health in the context of microtask crowdsourcing, and draw attention towards ensuring crowd worker wellbeing on such platforms. To the best of our knowledge, prior work has not explored the overall health of crowd workers in the microtasking ecosystem. It is known that health is a rather broad concept. In this work, our goal is to understand worker health from two main perspectives: 1) working ergonomics and physical health, and 2) psychosocial conditions and mental health.

As a precursor to developing methods and interventions to support worker health, we first explored means to improve worker health by explicitly asking crowd workers about their needs. Despite the abundance of healthcare applications, readily available on either desktop or mobile devices – including professional healthcare services like Teladoc[34], stress relief applications like Headspace[35], and simple break reminder such as AntiRSI[36] – many people remain unaware of the common health-related issues that can arise as a result of their work [232]. This particularly applies to crowd work, owing to the general lack of regulation and the relative recency of the online work paradigm. We envision that crowdsourcing platforms and task requesters can take measures to promote workers' health, and foster a sustainable relationship with crowd workers. For example, workers could regularly receive *health interventions* in the form of "tasks" that they are asked to complete, with an aim to maintain and improve their health, and increase an awareness of the potential health-related issues that they may encounter while completing crowdsourcing tasks. However, such interventions should first and foremost be informed by the workers' existing needs and their willingness to receive health-related interventions during work. It is important

[34] `https://www.teladochealth.com/`

[35] `https://www.headspace.com/`

[36] `http://antirsi.leverlabs.io/`



Andrea 15:01
In what mood are you today? Select the figure that best represents your current mood.

15:01
B

Andrea 15:01
Which of the following describes the income you earn from crowdsourced microtasks?

15:02
Secondary source of income

Andrea 15:02
How many hours do you work on crowdsourcing platform each day on average?

15:02
3-4

Andrea 15:02
Please indicate your usual working time on crowdsourcing platform in a day (24-hour system). Please select all the timeslots that apply.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AM | 0 - 1 | 1 - 2 | 2 - 3 | 3 - 4 | 4 - 5 | 5 - 6 | 6 - 7 | 7 - 8 | 8 - 9 | 9 - 10 | 10 - 11 | 11 - 12 |
| PM | 12 - 13 | 13 - 14 | 14 - 15 | 15 - 16 | 16 - 17 | 17 - 18 | 18 - 19 | 19 - 20 | 20 - 21 | 21 - 22 | 22 - 23 | 23 - 24 |

Figure 5.8: A screenshot of TickTalkTurk employed to gather survey responses from workers.

to understand what workers prefer in terms of the intervention types, their duration, and frequency. Therefore, we aim to address this through our first research question:

**RQ5.3:** What is the prevalent physical and mental health status of crowd workers in microtask crowdsourcing marketplaces?

**RQ5.4:** To what extent are healthcare interventions needed in crowdsourcing marketplaces? What are the preferred characteristics of such interventions from the perspective of workers?

To investigate these two research questions, we designed a survey consisting of 60 items related to 1) the basic demographics and working environment of crowd workers, 2) their working ergonomics and physical discomfort, 3) psychosocial conditions and mental health, and 4) worker needs. We employed TickTalkTurk (Figure 5.8) to gather survey responses from crowd workers, as in the previous chapters we have shown that conversational user interfaces are effective alternatives to traditional web-based survey as they can increase satisfaction and engagement. We conducted the study on two popular crowdsourcing platforms – Amazon Mechanical Turk (MTurk) and Prolific.

Results showed that workers across the two platforms faced similar health-related issues, but also differed from each other to some extent. In terms of their physical health, crowd workers across both platforms reported typically feeling less comfortable in their necks, shoulders and backs as a result of their work. This physical discomfort was found to be related to the working ergonomics of crowd workers; we found evidence from a correlation analysis suggesting that an unhealthy working posture was a likely cause for their physical discomfort. In terms of their mental health, we found that workers' energy levels (or fatigue) could be affected by task content. This was mediated by factors such as the meaning of work completed and possibilities of learning that their work offers. In addition, the mental wellbeing of crowd workers is affected by their work pace and task demands. Our findings suggest that platforms need to take a major responsibility, together with task requesters and academic researchers, in providing healthcare interventions around crowd work to improve worker health in crowdsourcing marketplaces.

This study has important implications in terms of understanding general worker health in current crowdsourcing platforms and informs the design of future methods to effectively improve worker health. In their influential work on charting out a future for crowd work several years ago, Kittur et al. [117] asked, "*Can we foresee a future crowd workplace in which we would want our children to participate?*". Through our study, we extend this conversation and argue that focusing on building methods and provisions to ensure worker wellbeing on crowdsourcing platforms is essential to make such a future possible.

## Related Work: Health of Crowd Workers

### Wellbeing of Crowd Workers

Previous work in the field of crowdsourcing studied worker health mainly from the perspectives of emotions and wellbeing. A recent technical report from Microsoft has comprehensively reviewed the past and envision the future of work [214]. The report emphasized the important role that wellbeing can play [26]. A recent study systematically review the relationship between the office working environment and employee health & wellbeing [36]. Furthermore, recent work has shown that the state-of-the-art methods can to some extent improve work productivity and wellbeing [230]. Another direction of relevant research relates to worker emotions and moods [48, 9, 223], since both emotion and mood are valenced affective responses. Prior studies have proposed a variety of instruments to measuring the emotion, such as the Self Assessment Manikin [21], the Affective Slider [11], the Achievement Emotions Questionnaire [166], and Pick-a-Mood (PAM) [48]. Based on the worker moods measured by PAM, researchers have presented the that crowd workers in a pleasant mood could be better engaged while completing online tasks, and meanwhile produce outcomes of higher quality with less cognitive taskload, either on traditional crowdsourcing platfroms [67, 235, 244] or in onversational crowdsourcing (Section 4.1).

Previous works have tangentially contributed towards improving worker wellbeing, by improving various factors that affect the dynamics of crowd work. Many prior studies have focused on invisible labor and power asymmetry, and proposed to build a healthy requester-worker relationship [190, 83, 66, 141]. Another popular research topic relates to improving workers' possibilities of development [13, 34, 208]. Atelier was therefore designed to re-purpose crowdsourcing tasks as mentored and paid skill development, named micro-internships [208]. Chiang et al. designed a system called Crowd Coach to assist workers in skill growth [34]. Others have proposed a variety of mechanisms to improve trust and ensure fair payment in crowdsourcing marketplaces [77, 229, 187, 192, 88]. Whiting et al. developed a tool for task requester to ensure the minimum wage [229]. Recent work by Savage et al. recommended transparency criteria to guide workers to earn higher salaries [192]. Furthermore, prior works attempted to improve worker wellbeing with a better task design, by improving task clarity [234, 138, 75, 108], and combining workers' opinions [23, 198].

In this study, we attempt to deeply investigate workers' health and wellbeing and try to understand the relationship between worker health and the working environment.

### Physical Discomforts Experienced by Workers

Physical discomforts and ergonomics of office work, particularly for sitting workers, have become an important research topic for decades [158]. Researchers have started to invent techniques and instruments to effectively assess working postures, and posture-related somatic problems [38, 82]. As technology advances, people tried to use more advanced and novel means, such as electromyography (EMG) measures [162], to measure body discomfort and assess its effect on work productivity. Recent studies started to focus on ergonomics and working postures while using computers [137, 177, 231]. In prior work by Luttmann et al., the authors performed precise measurements to assess muscular activities and working conditions [137]. An early study in 2011 specifically looked into body pain related to neck, shoulders, and arms, which are commonly complained about by computer office workers [177]. Woo et al. performed a systematic review to propose ergonomics standards and guidelines for computer workstation design, and summarize their effect on worker health [231]. In the context of crowdsourcing, researchers have found that both physical and mental fatigue could have negative impacts on crowd work [139, 242]. Other studies using mixed methods have revealed the diversity in the work environments at the disposal of crowd workers, and how these shape the quality of work that is produced [66]. Nevertheless, little is currently understood about the working ergonomics and physical health of crowd workers. We aim to address this important knowledge gap to inform future design choices that should focus on maintaining and improving the health of online crowd workers.

### Mental Health of Workers

Mental health has become a very important topic in society since it relates to everyone in the world [232]. There are many previous studies focusing on mental disorders and corresponding treatments (e.g. meditation, hypnosis, relaxation therapies, etc.) relating to office work [134, 224, 205, 212, 152]. To assess one's overall health and mental health, the SF-36 survey has been extensively used [227, 228], which has two subsets for evaluating mental health (i.e. mental wellbeing and work energy/fatigue). Another important aspect of mental health is workers' psychosocial conditions related to their overall working environments. A popular questionnaire for evaluating psychosocial work environment is the Copenhagen Psychosocial Questionnaire (COPSOQ) [120]. In the context of crowdsourcing, the cause of unhealthy psychosocial conditions and mental wellbeing could be power asymmetry and invisible labor [104, 190, 83, 141]. In crowdsourcing marketplaces, workers have to spend much time and invest much effort into underpaid tasks due to power asymmetry, meaning that task requesters usually have absolute power. Under this circumstance, Turkopticon was created [104] to enable workers to evaluate task requesters publicly. Some communities like Turkerview[37] or Reddit crowdsourcing-related groups[38] have emerged.

In addition to the prevailing platform dynamics in crowdsourcing marketplaces [189], content that workers consume as a result of accessing and completing on-demand work can also have a significant impact on their mental health and wellbeing. Long-term, continuous and extensive exposure to disturbing content has been found to have significant detrimental health consequences for people involved in such work [202, 167, 30, 79]. Recent work has shown that content moderation is a task prone to emotional exhaustion due to even relatively benign aspects such as the incivility of the content [181, 204]. Research has shown that

[37] https://turkerview.com/

[38] https://www.reddit.com/r/mturk, https://www.reddit.com/r/ProlificAc/

content moderators are regularly exposed to far more malignant content [31]. To address this problem, researchers have used blurring to reduce the harmful content exposure time for moderators [44].

In this study, we will assess workers' mental health and psychosocial working environment to develop an understanding of the prevailing worker health on crowdsourcing platforms and better inform health related interventions for the future.

### Survey Design

We designed a survey consisting of 60 items, delivered through a conversational user interface. The items in the survey relate to 1) the basic demographics and working environment of crowd workers, 2) their working ergonomics and physical discomfort, 3) psychosocial conditions and mental health, and 4) worker needs.

#### Using Conversational Crowdsourcing for Survey Completion

Conversational user interfaces have become increasingly common as an alternative to the traditional graphical user interfaces. In Chapter 3, we have shown that a conversational interface or a chatbot is capable of providing better user engagement and satisfaction. In addition, it has the potential of leading to a better output quality [113], due to its human-like means of interaction. Conversational interfaces have been successfully employed in a variety of domains [238], ranging from design [222] to search [6, 115, 226]. In the field of healthcare, chatbots and conversational interfaces have also been successfully used to play the role of an assistant for either patients or therapists [126, 150, 178], and for mental-health support or treatment [63, 209, 168, 135, 1].

In this study we employed TickTalkTurk to guide workers to complete the survey, instead of using traditional web-based survey forms. A screenshot of the conversational interface for completing the survey is shown in Figure 5.8. Conversational elements such as greetings, the response delay, and repeating the worker response are applied on the chatbot to improve conversational experience. On successfully completing the survey, the conversational agent provides workers with a survey completion code, which they can use to earn their rewards.

#### A Questionnaire for Understanding Worker Health

***Part I: Worker Background.*** In the first part of the questionnaire, we ask background questions to understand the demographic information and working environment of crowd workers. As listed in Table 5.3, the first part contains 14 questions. The first three questions pertain to the basic background in terms of gender, age, and current mood of workers. Next, we use 4 questions (4-7) to gain insights into the context of their participation on crowdsourcing platforms, about workers' income from crowdsourcing platforms (i.e. primary or secondary source of income), working hours per day on average, typical working time, and crowd work experience (in years).

Next, we aimed to understand workers' general working environment and immediate working environment [158] using questions 8-11. The *general working environment* refers to external elements such as lighting, temperature, humidity, noise, etc. The *immediate working environment* refers to working devices and setups, in terms of control, display, compatibility, layout, posture, etc. Note that in the survey, we did not use the terms "general working environment" and "immediate working environment" since interpreting these terms

Table 5.3: The questions used in the first part of the questionnaire: worker demographics and background.

| No. | Question | Answer type |
|---|---|---|
| 1 | May I know you gender? | Single-selection |
| 2 | How old are you? | Single-selection |
| 3 | In what mood are you today? | Single-selection |
| 4 | Which of the following describes the income you earn from crowdsourced microtasks? | Single-selection |
| 5 | How many hours do you work on MTurk/Prolific each day on average? | Single-selection |
| 6 | Please indicate your usual working time on MTurk/Prolific in a day. | Multiple-selection |
| 7 | For how long have you been working on MTurk/Prolific? | Single-selection |
| 8 | To what extent do you think your current working environment is comfortable, in terms of lighting, temperature, humidity, noise, etc. | 7-pt Likert-scale |
| 9 | So your current working environment is comfortable/uncomfortable, then do you think it is healthy? | 7-pt Likert-scale |
| 10 | To what extent do you think your current working setup and devices are comfortable, in terms of control, display, compatibility, layout, posture, etc. | 7-pt Likert-scale |
| 11 | So your current working setup and devices are is comfortable/uncomfortable, then do you think it is healthy? | 7-pt Likert-scale |
| 12 | Do you consider that you have colleagues (eg. other crowd workers)? | Single-selection |
| 13 | Do you share workspaces with your colleagues or work together in a shared work environment? | 5-pt Likert-scale |
| 14 | Do you take some measures to keep yourself healthy? (If so, what do you do?) | Free-text |

accurately requires specific domain knowledge. Instead, we simply used “working environment” and “working setup and devices” in the questions, to ensure accurate interpretation by workers [66]. For both general and immediate working environments, we first asked workers about their perceived degree of comfort, and then asked them about their perceived degree of health, to see whether workers can discern the difference between being comfortable and being healthy. These four questions are required to be answered using 7-point Likert-scales (from ‘*1: Very Uncomfortable*’ to ‘*7: Very Comfortable*’, or from ‘*1: Very Unhealthy*’ to ‘*7:Very Healthy*’). In terms of questions 9 and 11, before asking the perceived degree of health, the chatbot repeats the answers of questions 8 and 10 respectively, since previous work has shown that the use of repeating strengthens the involvement of conversation [211]. We used two questions to investigate whether workers perceive other crowd workers as being colleagues, and whether they work alone in their workspaces, because loneliness has been proved to have deleterious effects on health [27]. Finally, we asked workers if they take any measures to keep themselves healthy.

***Part II: Working Ergonomics and Physical Health.*** The second part of our questionnaire addresses workers’ working ergonomics and physical health status. Literature offers an abundance of instruments and tools to measure working ergonomics [38, 82, 231]. However, these were designed to specifically cater for traditional office work, or require precise measurement of distances, lengths, and angles. Therefore, based on the prior studies on ergonomics [82, 137, 231] and Stanford’s computer workstation ergonomics self-evaluation form [58], we designed a questionnaire that covers the relevant aspects of working ergonomics — chair, keyboard and mouse, screen/monitor, breaks/practices, and overall posture (question 15-23 in Table 5.4). Questions in all aspects were selected from validated surveys used in previous studies [82, 58, 137, 231].

In terms of the overall posture, we first asked workers about their working postures (sitting, standing, or other postures). Using example pictures of healthy sitting/standing

Table 5.4: The questions used in the second part of the questionnaire: working ergonomics and physical health.

| No. | Question | Answer type |
|---|---|---|
| 15 | What is your primary working posture? | Single-selection |
| 16 | Looking at these examples of healthy working postures, to what extent do you think your working posture is healthy? (showing Figure 5.9 of proper sitting and standing postures) | 7-pt Likert-scale |
| 17 | *If the posture includes sitting*: How often do you use armrests? | 5-pt Likert-scale |
| 18 | *If the posture includes sitting*: Can you indicate your sitting position? | Single-selection |
| 19 | *If the posture includes sitting*: How do you use your backrest? | 5-pt Likert-scale |
| 20 | How often do you take a break? | Single-selection |
| 21 | What is the distance between you and your screen? | 5-pt Likert-scale |
| 22 | Can you indicate the position of the top of your screen? | 5-pt Likert-scale |
| 23 | Can you indicate your keyboard/mouse placement? | Multiple-selection |
| Please tell me how comfortable your different body parts feel on an average day working on MTurk/Prolific. | | |
| 24 | Your eyes? | 7-pt Likert-scale |
| 25 | What about your head? | 7-pt Likert-scale |
| 26 | And your neck and shoulders | 7-pt Likert-scale |
| 27 | How is your back | 7-pt Likert-scale |
| 28 | What about your seat and thighs | 7-pt Likert-scale |
| 29 | And your knees and feet | 7-pt Likert-scale |

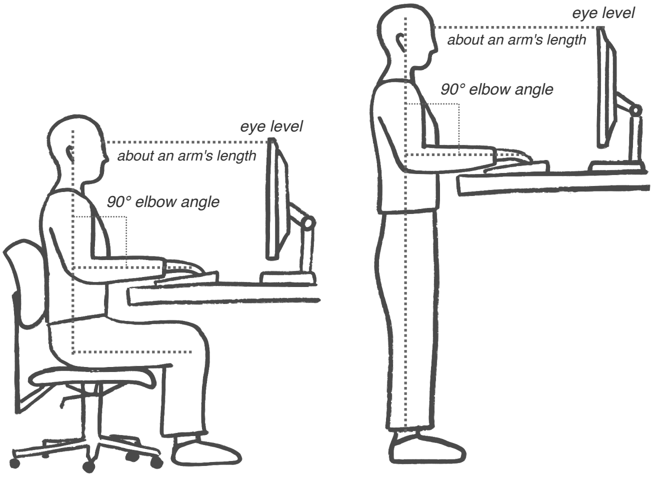


Figure 5.9: The proper sitting and standing postures of crowd work.

postures (Figure 5.9), we asked workers to rate the degree of health of their overall postures using a 7-point likert-scale (ranging from '*1: Very Unhealthy*' to '*7: Very Healthy*'). If the primary working posture of workers was found to be '`sitting`', we followed-up with three additional questions about chair settings (question 17-19). Workers are asked to report their frequency of using chair armrests, their positions as they sit on the chair ('`front edge`', '`middle`', or '`back`'), and the frequency of using chair backrest. We also ask workers how often they take a break.

As for the screen position, we gathered information about 1) the distance between the worker and the screen, and 2) the vertical position of the screen top. As prescribed by recent work [231], the position of the screen is deemed to be healthy if the distance to the worker is about an arm's length, and the screen top is at the eye level. Furthermore, workers were asked to report their keyboard/mouse positions. Their positioning is considered to be healthy if the worker can easily reach the keyboard/mouse while maintaining an elbow angle of 90 degrees (as shown in Figure 5.9).

According to the categories of different body parts used in ergonomics research [82], we asked workers to rate their perceived degree of comfort (on an average working day) with respect to each body part. Apart from the body parts mentioned in previous work, we added a question exploring the degree of comfort perceived with respect to workers' eyes. Since the nature of microtask crowdsourcing implies spending large amounts of time looking at screens, we believe this to be of important relevance. Workers were asked to answer questions about physical discomfort (questions 24-29) using 7-point Likert-scales indicating their perceived degree of comfort (ranging from '*1: Very Uncomfortable*' to '*7: Very Comfortable*').

***Part III: Psychosocial Condition and Mental Health.*** In Part III, we are interested in investigating the psychosocial condition and mental health of crowd workers. Psychosocial working conditions and working environments refer to working situation, work methods and pace, understanding of work process, possibilities of development, human-contacts and cooperation for work, etc. [78, 119]. In this study, we used subsets of existing instruments to measure pyschosocial conditions, mental wellbeing, and working energy/fatigue [120, 227, 228] as shown in Table 5.5.

In order to measure psychosocial conditions comprehensively, at least 44 questions are required (and proved to be valid) in a short version of COPSOQ [120, 227, 228]. Nevertheless, we did not use the a completely valid subset of COPSOQ as it would make the size of our survey too large. For the purpose and scope of this study, we reason this to be an acceptable trade-off, owing to our focus on overall worker health rather than systematically analyzing workers' psychosocial conditions. We selected 10 representative questions from the COPSOQ CORE items to address 9 representative dimensions that relate to crowd work — quantitative demands, work pace, emotional demands, influence at work, possibility for development, meaning of work, social support (supervisor and co-worker), feedback at work, and sense of community, belonging to three categories (type of production an tasks, work organization and job content, and interpersonal relations (as shown in Table 5.5). Questions were slightly reformulated to adapt them to the context of online crowdsourcing. In particular, *possibility for development* was adapted to represent the possibility of learning new things during crowd work instead of career promotion, as it is known that current crowdsourcing marketplaces lack career ladders [117]. *Social support from supervisor* was adapted to refer to the help and support from crowdsourcing platforms and task requesters, while *sense of community* captures the extent to which workers are aware of worker forums and unities. Responses to the questions about pyschosocial conditions were gathered using 5-point Likert-scales of either frequency ('*1: Never*' to '*5: Always*') or intensity ('*1: To a Very Small Extent* ' to '*5: To a Very Large Extent*'), as recommended by previous work [120].

To measure worker mental health, in total 10 questions were selected from SF-36 [227, 228]. One question is for self-reporting general health (poor, fair, good, very good, and excellent). Furthermore, we used two validated subsets (the other 9 questions) for measuring mental wellbeing and working energy/fatigue from SF-36, where 4 questions are used for measuring working energy and fatigue and 5 questions are used for measuring emotional wellbeing. The final emotional wellbeing score or energy/fatigue score is the average of the scores of all questions in the corresponding dimension.

***Part IV: Workers' Needs.*** The last part of the questionnaire is devoted to the inquiry

Table 5.5: The questions used in the third part of the questionnaire: psychosocial condition and mental health. (R) represents that the final score should be reversed.

| No. | Question | Dimension | Answer type |
|---|---|---|---|
| Type of production and tasks | | | |
| 30 | How often do you have enough time for tasks on MTurk/Prolific? | Quantitative demands | 5-pt Likert-scale |
| 31 | Do you have to work very fast? | Work pace (R) | 5-pt Likert-scale |
| 32 | Is completing tasks on MTurk/Prolific emotionally demanding? | Emotional demands (R) | 5-pt Likert-scale |
| Work organization and job content | | | |
| 33 | Do you have a large degree of influence on the decisions concerning completing tasks on MTurk/Prolific? | Influence at work | 5-pt Likert-scale |
| 34 | Do you have the possibility of learning new things through completing tasks on MTurk/Prolific? | Possibilities for development | 5-pt Likert-scale |
| 35 | Do you feel that completing tasks on MTurk/Prolific is meaningful? | Meaning of work | 5-pt Likert-scale |
| Interpersonal relations | | | |
| 36 | How often do you get help and support from MTurk/Prolific or task requesters, if needed? | Social support (supervisor) | 5-pt Likert-scale |
| 37 | How often do you get help and support from other workers, if needed? | Social support (co-worker) | 5-pt Likert-scale |
| 38 | How often do task requesters bonus/message you because how well you carry out your work? | Feedback at work | 5-pt Likert-scale |
| 39 | Is there a good atmosphere between you and other workers (on either crowdsourcing platforms or other worker forums e.g. Reddit)? | Sense of community | 5-pt Likert-scale |
| 40 | In general, would you say your health is excellent, very good, good, fair or poor? | General health | Single-selection |
| 41 | While completing tasks on MTurk/Prolific, do you feel full of pep? | Fatigue/energy (R) | 6-pt Likert-scale |
| 42 | While completing tasks on MTurk/Prolific, have you been a very nervous person? | Emotional well-being | 6-pt Likert-scale |
| 43 | While completing tasks on MTurk/Prolific, have you felt so down in the dumps that nothing could cheer you up? | Emotional well-being | 6-pt Likert-scale |
| 44 | While completing tasks on MTurk/Prolific, have you felt calm and peaceful? | Emotional well-being (R) | 6-pt Likert-scale |
| 45 | While completing tasks on MTurk/Prolific, do you have a lot of energy? | Fatigue/energy (R) | 6-pt Likert-scale |
| 46 | While completing tasks on MTurk/Prolific, have you felt downhearted and blue? | Emotional well-being | 6-pt Likert-scale |
| 47 | While completing tasks on MTurk/Prolific, do you feel worn out? | Fatigue/energy | 6-pt Likert-scale |
| 48 | While completing tasks on MTurk/Prolific, have you been a happy person? | Emotional well-being (R) | 6-pt Likert-scale |
| 49 | While completing tasks on MTurk/Prolific, do you feel tired? | Fatigue/energy | 6-pt Likert-scale |

of workers' needs. Based on the results of this study, we aim to draw attention to crowdsourcing platforms and task requesters to the fact that the health of crowd workers should fundamentally matter. We hope to inform future measures, policy decisions, or interventions that put workers' health at the forefront of design choices. Therefore, we used 10 questions to elicit workers' needs and acquire an understanding of workers' perspective on this matter. At the end of the survey, workers can optionally provide any further comments, remarks or suggestions. The questions of Part IV are listed in Table 5.6.

The first two questions are about workers' needs with regard to physical health and mental health respectively. To improve worker health, we hypothesise that workers could regularly receive interventions (breaks/exercises/treatments) while completing crowdsourcing tasks, which can help them at least be aware of the potential health-related problems

Table 5.6: The questions used in the fourth part of the questionnaire: workers' needs.

| No. | Question | Answer type |
|---|---|---|
| 50 | For which part(s) of your body do you think you need some physical exercises? | Multiple-selection |
| 51 | For which aspect(s) of your psychosocial condition do you think you need improvements? | Multiple-selection |
| 52 | To what extent will you be happy to use a tool that provides breaks/exercises/treatments to improve your overall health while completing crowdsourcing tasks? *Optional: Can you tell me why?* | 5-pt Likert-scale & Free-text |
| 53 | What features would you like to see in such a tool, considering that they are all backed by scientific evidence? *Optional: Can you tell me why?* | Multiple-selection & Free-text |
| 54 | What type of working modes of this tool would you prefer? *Optional: Can you tell me why?* | Multiple-selection & Free-text |
| 55 | Do you think that you should get paid while you are using the tool to take some breaks/exercises/treatments? *Optional: Can you tell me why?* | Single-selection & Free-text |
| 56 | How would you like to receive interventions (breaks/exercises/treatments)? | Single-selection |
| 57 | How long would you like the interventions (breaks/exercises/treatments) from the tool to be? | Single-selection |
| 58 | How frequently would you like to take breaks/exercises/treatments from such a tool? | Single-selection |
| 59 | Who do you think should be responsible for developing the tool? Please check all that apply. | Multiple-selection |
| 60 | Do you have any other comments, remarks, or suggestions? Your thoughts are valuable to us. | Free-text |

they may encounter in crowd work and contribute towards alleviating them. With questions 52-55, we first asked workers to what extent they would be happy to receive interventions, and then asked questions about their preferred features ('`simple breaks`', '`physical exercises`', or '`treatments for mental health`'), preferred working mode ('`pull`' or '`push`', i.e. asking for interventions when they want to, or enabling a tool to actively sending interventions to them), and whether they would like to get paid while taking interventions. For these four questions, workers can also provide free comment to explain their answers. The remaining questions are about how they would like to receive interventions ('`between task batches`', '`within task batches`', or '`outside of the platform`'), the size of the intervention (length in minutes), the frequency of the interventions, and whom they believe this responsibility lies with ('`crowdsourcing platforms`', '`task requesters`', '`academic researchers`', '`crowd workers`', or '`third-parties`').

## Study Setup

### Platforms and Settings

We conducted the study on two popular crowdsourcing platforms — Amazon Mechanical Turk (MTurk) and Prolific, featuring different usage profiles [165]. MTurk has been designed, and is commonly used for data labeling tasks, where a human intelligence task (HIT) usually represents a microtask asking workers to annotate images, transcribe audios, analyze the sentiment of text, etc. [51]. MTurk workers tend to perform repetitive tasks within large batches [49]. On the other hand, Prolific has been designed to serve research study needs. It is commonly used for carrying out user studies in the realms of behavior research, user and market research.

On each crowdsourcing platform, we recruited 150 workers. Considering that workers come from all over the world and work in different time zones, we published surveys in three batches throughout a day. This means that for each platform, we published 50 tasks every 8 hours. Note that the purpose of this study was not to explore the differences between worker health across different demographics, but to build a general perspective of the status quo. Therefore, we did not enforce specific restrictions in participation. Future work could systematically address and model similarities and differences in workers' health based on demographic features. This study has been approved by the human research ethics committee of our institute.

### Quality Control

Our broad goal is to understand the general health of crowd workers. Therefore, we did not set any qualifications to pre-screen workers. To improve the quality of the overall analysis, surveys included three attention check questions [140]. The attention check questions is rather simple: "It is important that you pay attention to this study. Please select `'Strongly Agree'`", and workers are supposed to select the correct answer from five options in total. Workers who fail any of the three attention check questions are excluded from analysis. Workers were compensated regardless of their success in passing the attention check questions.

### Compensation

Through a pilot run, we estimated a survey completion time of 15 minutes (800 seconds) and initially paid workers USD 2.5 (or GBP 1.88) per task. We found that the average survey completion time across two platforms was $850.28 \pm 324.35$ (seconds); $961.43 \pm 334.29$ (seconds) on Prolific, and $728.43 \pm 263.89$ (seconds) on MTurk. To ensure fair pay, we granted bonuses to workers whose active task execution time was longer than 15 minutes (in total GBP 86.6 for bonusing Prolific workers, and USD 41.2 for bonusing MTurk workers), resulting in the actual average hourly wage of USD 12.8. In total, we paid GBP 368.6 and USD 416.2 to gather responses from 300 workers across the two crowdsourcing platforms.

## Results

On Prolific, all the workers completing the survey will give the same completion code by default. We applied the same working mode on MTurk. However, we had to exclude 7 submissions from MTurk because of re-using the completion code for multiple MTurk accounts. Furthermore, for analysis, we excluded 2 workers from Prolific because at least one (out of three) attention check question was not correctly answered. On MTurk, 8 workers were excluded due to the same reason. As a result, 148 Prolific workers and and 135 MTurk workers were included in our analysis.

### Demographics and Background

The majority of participating crowd workers were male. On Prolific, of 148 valid submissions, 90 were male; 57 were female; and one worker reported non-binary. On MTurk, the gender distribution was similar, where 92 workers were male and 43 workers were female.

Figure 5.10 shows the workers' country of residence. Prolific users came from Europe, North America, South America, Africa, and Asia. The majority of the workers came from Europe (35 from the United Kingdom; 31 from Portugal; 15 from Poland; and other 13 European countries). There were 20 workers from North America (Mexico, US, and Canada), 8 workers from South America (Chile), 5 workers from Africa (South Africa), and 2 workers from Asia (Korea and Israel). On MTurk, 93 out of 135 workers reported to be residing

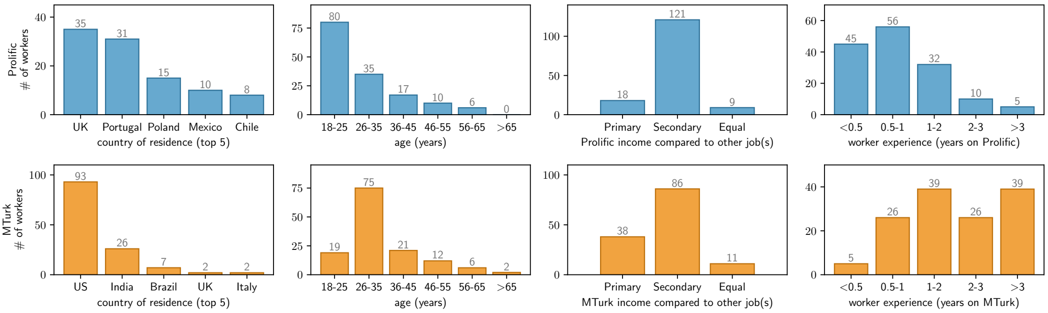


Figure 5.10: Worker demographics in terms of the country of residence, age, income, and working time on two crowdsourcing platforms (Prolific and MTurk).

in the United States; 26 workers reported to be living in India; 8 workers reported South America (Brazil and Peru); the rest came from Europe (5 countries).

As shown in Figure 5.10, the workers on Prolific were generally younger (80 workers - out of 148 - reported that they were younger than 25 years old) while on MTurk the majority (75 out of 135) of the workers were 26-35 years old. Moreover, the Prolific workers were also less experienced compared to the MTurk workers, since 101 (68%) Prolific workers had been working on the platform less than one year, while the MTurk workers were more experienced since 104 (77%) of them had been working on MTurk over 1 year (particularly, 39 workers reported that they had been working on MTurk longer than 3 years). In terms of the source of income, on both Prolific and MTurk, most workers used the crowdsourcing platform as a secondary income source (121 and 86 on Prolific and MTurk respectively), but more workers earned their primary income from MTurk (38) compared to Prolific (18).

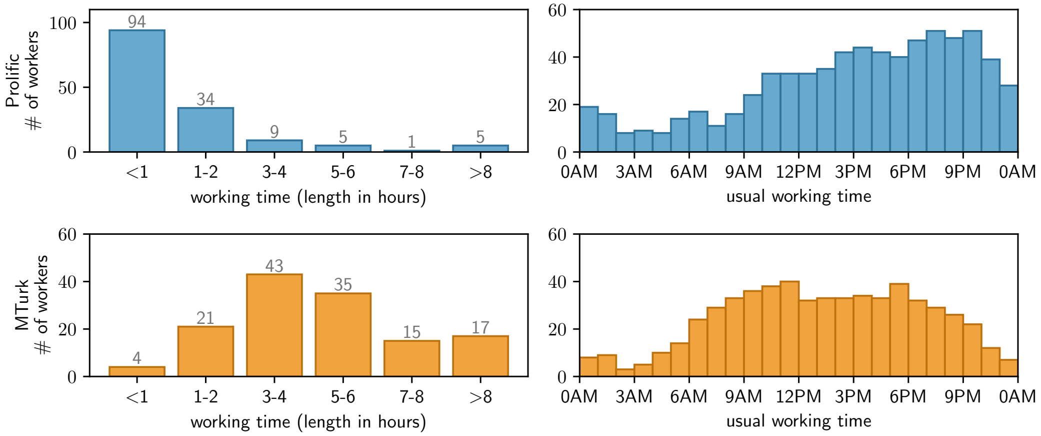


Figure 5.11: Workers' usual working time throughout a day on two crowdsourcing platforms (Prolific and MTurk).

Investigating workers' usual working time is an important part of understanding worker health. As shown in Figure 5.11, we can observe dramatic difference between the Prolific workers and the MTurk workers. Of 148 valid Prolific workers, 94 (63%) reported that they worked less than 1 hour per day, and only 20 (14%) workers worked longer than 3 hours on average. In addition, the majority of them liked to work in the afternoon and in the evening, according to Figure 5.11. However, as for MTurk, 81% (110 out of 135) of the workers spent longer than 3 hours per day on completing crowdsourcing tasks. Also, the MTurk workers tended to work in the morning and afternoon. While striving to minimise the effect of task distribution time on the collected results, (we published surveys evenly by three batches throughout a day), we acknolwedge that results could be affected by task publishing time.

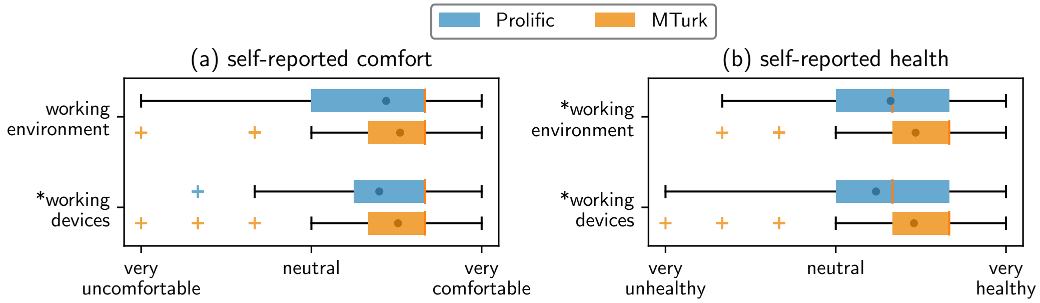


Figure 5.12: Self-reported (a) comfort and (b) health in terms of working environment (lighting, temperature, humidity, noise, etc.) and working devices (control, display, compatibility, layout, posture, etc.) on two crowdsourcing platforms (Prolific and MTurk). Asterisk (*) represents significance difference between Prolific workers and MTurk workers.

Furthermore, to have a better understanding of workers' working environment, we asked whether workers felt comfortable and healthy about their general (lighting, temperature, humidity, noise, etc.) and immediate (devices and setups, in terms of control, display, compatibility, layout, posture, etc.) working environments [158]. As shown in Figure 5.12, Prolific workers reported less comfort and health scores compared to the MTurk workers. Since self-reported comfort scores and self-reported health scores do not come from normal distributions (Shapiro-Wilk tests $p < 0.05$ for all groups), we conducted Mann-Whitney U tests to test the significance. We found that Prolific workers' self-reported scores of *comfort of working devices*, *health of working devices*, and *health of working environment* are significantly lower than MTurk workers ($p = 0.015$, $p < 0.001$, and $p = 0.003$ respectively). In addition, we found that on both platforms self-reported health scores (Prolific: $4.84 \pm 1.36$, MTurk: $5.39 \pm 1.26$) are lower than self-reported comfort scores (Prolific: $5.26 \pm 1.40$, MTurk: $5.54 \pm 1.28$), which implies workers can possibly discern differences between comfort and health and may realize that their working environments are unhealthy although they feel comfortable. However, this finding is inconclusive since health scores and comfort scores are measured using different metrics, so they are not supposed to be compared statistically.

On Prolific, 97 workers (66%) did not have colleagues (co-workers), and 84 workers (57%) worked alone (never not share a workspace with others). On MTurk, the situation is slightly different, where 54 workers (40%) claimed that they had no colleagues and 66 workers (49%) never shared a workspace with others. This, all in all, suggested that a large number of workers are working alone (either had no co-workers or never shared a workspace with others) on crowdsourcing platforms.

***Summary of background analysis:***

1. Workers on MTurk were, on average, older and longer engaged in microtask crowd work on the platform than Prolific workers.

2. Workers on MTurk tended to work longer and usually worked in the morning and afternoon, while Prolific workers worked less and usually worked in the afternoon and evening.

3. According to self-report, MTurk workers had better working environments compared to Prolific workers.

4. On both platforms, a large proportion of workers worked alone (no colleagues or no shared workspace, particularly on Prolific).

### Ergonomics and Physical Health

In the second part of the survey, we investigated the working ergonomics and physical health of the workers, and the correlation between working ergonomics and physical health.

A proper working posture is essential to one's health. As shown in Figure 5.13 (a), the distributions of working postures are similar across two platforms. The majority of workers on Prolific (116) and MTurk (94) were sitting. In addition, 22 Prolific workers and 30 MTurk workers could both sit and stand while completing tasks. Moreover, we showed example pictures of proper working postures (Figure 5.9) to workers and asked them to rate the degree of health of their overall working postures by comparing the examples with their own working postures. Results revealed that, as shown in Figure 5.13 (b), most MTurk workers (93 out of 135) reported that their working postures should belong to the categories of healthy, while 27 MTurk workers believed that their working postures were unhealthy. In contrast, the Prolific workers' postures were relatively unhealthier in comparison with the MTurk workers. In terms of frequency of breaks, workers from both platforms shared a similar pattern – most of them did not take breaks too frequently (more often than every 30 minutes), and actively took breaks at least every 4 hours, as can be seen in Figure 5.13 (c).

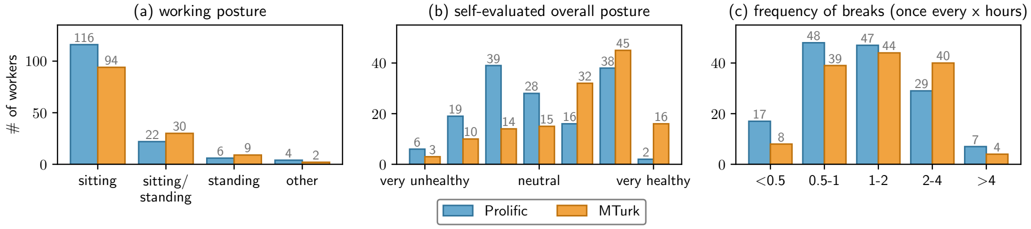


Figure 5.13: Self-evaluated working postures and physical health across two crowdsourcing platforms (Prolific and MTurk).

In terms of the use of armrests, as shown in Figure 5.14 (a), most MTurk workers had armrests (only 14 did not have) and would like to use them (only 21 reported that they rarely used them). However, as for Prolific, a considerable proportion of workers (30%) did not have armrests at all, and the workers who had them also used them less frequently in comparison with the MTurk workers. According to Figure 5.14 (b) and (c), more Prolific workers were used to sit to the back of the chair and lean on the backrest. Nevertheless, more MTurk workers liked to sit upright with the support of the backrest, which is proved to be a healthier working posture [231].

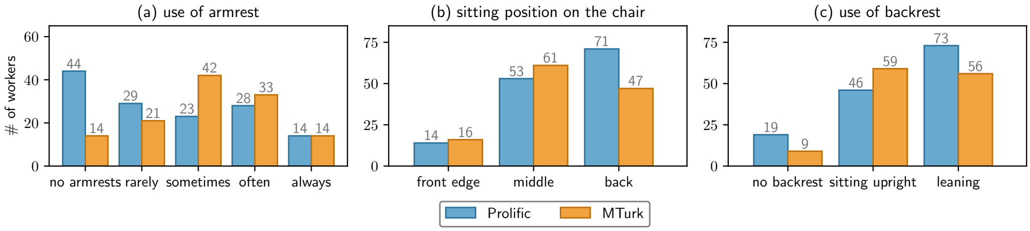


Figure 5.14: Self-evaluated sitting postures across two crowdsourcing platforms (Prolific and MTurk).

As for working ergonomics, we finally asked workers about their devices (screen, keyboard, and mouse). The workers from both Prolific and MTurk, in general, reported that the distance to the screen was around an arm's length, as shown in Figure 5.15 (a). In comparison with the MTurk workers, the tops of more Prolific workers' screens were below their eye levels, whereas previous work considered that the screen top being at the eye level is healthy [231]. Moreover, we found that many workers (42 on Prolific and 47 on MTurk) reported their screen tops were higher than the eye level. This is possibly due to the large screen size. With regard to the keyboard and the mouse (Figure 5.15 (c)), we found that only 25 workers (11 from Prolific and 14 from MTurk) had to overreach their shoulders and arms in order to use the keyboard/mouse. It is also worth mentioning that 60 workers (44%) from MTurk reported that the position of the keyboard/mouse supported a 90-degree elbow angle, which is another health working setup according to the previous study [231]. The results above suggest that, on average, MTurk workers have healthier working postures and working setups, compared to Prolific workers.

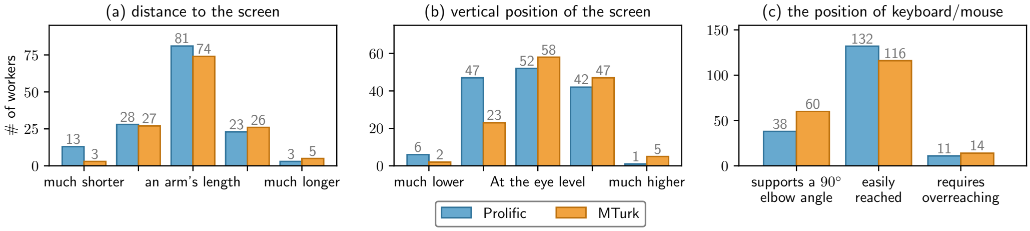


Figure 5.15: Self-evaluated positions of working devices across two crowdsourcing platforms (Prolific and MTurk).

Apart from working ergonomics and setups, workers were asked to report how comfortable their different body parts feel on an average day working on the crowdsourcing platform. Results are shown in Figure 5.16. Since the comfort scores of all body parts on both platforms do not come from normal distributions (Shapiro-Wilk tests $p < 0.05$ for all groups),

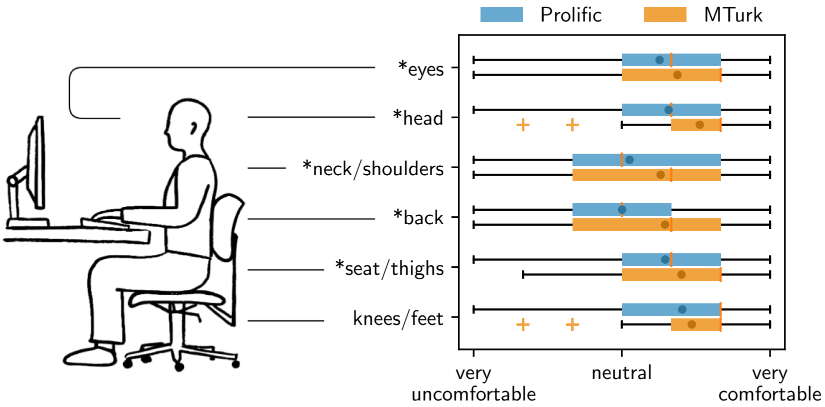


Figure 5.16: Self-reported discomfort scores of body parts across two crowdsourcing platforms (Prolific and MTurk). Asterisk (*) represents significance difference between Prolific workers and MTurk workers.

we conducted Kruskal-Wallis H-test to test the null hypothesis that median comfort scores of all of the body parts are equal. Significance testings suggested that for both Prolific and MTurk ($H = 70.35, p < 0.001$ and $H = 20.49, p = 0.001$ respectively), median comfort scores of different body parts were not equal. As we can see from Figure 5.16, the neck/shoulders and the back are the body parts that mainly make workers uncomfortable on an average working day. We also found that the physical discomfort problems were more serious on Prolific workers' bodies across all the body parts except knees/feet (significant differences found in eyes, head, neck/shoulders, back, and seat/thighs tested by Mann-Whitney U tests, $p < 0.02$ for all groups). Clearly, the result of physical discomfort is aligned with the condition of workers' postures and setups, as the MTurk workers in general had healthier working postures/setups which results in less physical discomfort.

To have a better understanding of whether healthier working posture and setup can result in less physical discomfort, we conducted correlation analysis between self-reported evaluation of working ergonomics (overall posture, screen distance, position, etc.) and body discomfort. We selected five features of working ergonomics that can be ranked (by assigning numerical scores to the answers, which are proportional to the degree of health) – the overall working posture shown in Figure 5.13 (b), the frequency of breaks shown in Figure 5.13 (c), the frequency of using armrest shown in Figure 5.14 (a), the distance between the worker and the screen shown in Figure 5.15 (a), and the vertical position of the screen shown in Figure 5.15 (b). Note that for the screen distance, "an arm's length" has the highest score, while both "much shorter" and "much longer" have the lowest score. This also applies to the screen's vertical position. The results of Spearman's correlation analysis (the correlation coefficients $\rho$ and $p$-values) are reported in Table 5.7. The overall posture significantly relates to the degree of discomfort of all the body parts, particularly to back ($\rho > 0.5$), neck/shoulders, eyes, and head ($\rho > 0.4$). The correlation coefficients $\rho$ values of the other relations (break frequency, screen position, and use of armrest) are relatively low ($\rho < 0.2$), implying that body discomfort is possibly caused by the overall posture and synthetic factors, rather than some specific ergonomics features.

Table 5.7: The Spearman's rank correlation coefficients ($\rho$) of relations between the self-evaluated postures (only variables that can be ranked were selected) and the self-reported body discomforts. Asterisk (*) represents significance.

| | Body discomfort | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Eyes | | Head | | Neck/shoulders | | Back | | Seat/thighs | | Knees/feet | |
| | $\rho$ | $p$-value | $\rho$ | $p$-value | $\rho$ | $p$-value | $\rho$ | $p$-value | $\rho$ | $p$-value | $\rho$ | $p$-value |
| ***Overall posture*** | 0.451 | <0.001* | 0.425 | <0.001* | 0.477 | <0.001* | 0.523 | <0.001* | 0.376 | <0.001* | 0.325 | <0.001* |
| ***Frequency of breaks*** | 0.042 | 0.477 | 0.067 | 0.261 | 0.035 | 0.563 | 0.023 | 0.695 | 0.069 | 0.246 | 0.064 | 0.286 |
| ***Screen distance*** | 0.093 | 0.119 | 0.125 | 0.036* | 0.098 | 0.099 | 0.081 | 0.174 | 0.091 | 0.129 | 0.145 | 0.015* |
| ***Screen position*** | 0.100 | 0.093 | 0.106 | 0.076 | 0.115 | 0.052 | 0.126 | 0.034* | 0.105 | 0.078 | 0.027 | 0.647 |
| ***Use of armrest*** | 0.095 | 0.172 | 0.157 | 0.023* | 0.129 | 0.063 | 0.191 | 0.005* | 0.121 | 0.081 | -0.122 | 0.078 |

***Summary of physical health analysis:***

1. MTurk workers generally had healthier working postures/setups than Prolific workers.
2. Workers on both platforms reported that they felt less comfortable in their necks, shoulders, and backs. Prolific workers physically felt less comfortable in general, compared to MTurk workers.
3. The working posture significantly relates to the body discomfort.

### Psychosocial Condition and Mental Health

In the third part of the survey, we focus on the psychosocial condition and mental health, and their relations. Results are reported in Table 5.8. The score ranges from 1 to 5. The higher the score is, the better the psychosocial condition could be. For the dimension of work pace and emotional demands, noted with "(R)", their scores have been reversed, because the questions in these two dimensions are worded negatively while the other questions are worded positively.

Psychosocial scores of all dimensions on both platforms do not come from normal distributions (Shapiro-Wilk tests $p < 0.05$ for all groups). We therefore performed Mann-Whitney U tests to find differences between the Prolific workers and the MTurk workers. Significant differences were found in the dimensions of work pace ($p < 0.001, CL = 0.33$, CL means the common language effect size thereafter), emotional demands ($p < 0.001, CL = 0.27$), meaning of work ($p = 0.015, CL = 0.43$), social support from requesters ($p = 0.013, CL = 0.42$), social support from colleagues ($p < 0.001, CL = 0.27$), feedback at work ($p < 0.001, CL = 0.27$), and sense of community ($p = 0.005, CL = 0.42$). Results shows that the workers on MTurk generally had to work faster, and they considered the tasks on MTurk to be more emotionally demanding, compared to the Prolific workers. The Prolific workers believed that their crowd work was more meaningful. However, the MTurk workers did get more help and support from the platform, requesters, and other workers. The results also showed that the MTurk workers received bonus and positive feedback more frequently. In summary, the Prolific workers exhibited better pychosocial conditions in terms of type of production/task (work pace, and emotional demands) and work content (meaning of work); the MTurk workers exhibited better psychosocial conditions in interpersonal relations (social support, feedback at work, and sense of community).

The health scores of mental wellbeing and energy/fatigue are displayed as boxplots in Figure 5.17. The health score ranges from 0 to 100. The higer the score is, the healthier the worker is. Interestingly, in terms of mental health, we found completely opposite

Table 5.8: Self-reported psychosocial scores of workers on Prolific and MTurk platforms. Asterisk (*) represents significance difference between Prolific workers and MTurk workers.

| ***Dimension*** | **Question** | **Scale (1-5)** | **Prolific** | **MTurk** |
|---|---|---|---|---|
| *Quantitative demands* | How often do you have enough time for tasks on MTurk/Prolific? | Never - Always | 3.82 ± 0.85 | 3.90 ± 0.75 |
| *Work pace (R)** | Do you have to work very fast? | To a very small extent - To a very large extent | 3.58 ± 0.99 | 2.89 ± 1.14 |
| *Emotional demands (R)** | Is completing tasks on MTurk/Prolific emotionally demanding? | To a very small extent - To a very large extent | 4.22 ± 0.89 | 3.24 ± 1.23 |
| *Influence at work* | Do you have a large degree of influence on the decisions concerning completing tasks on MTurk/Prolific? | To a very small extent - To a very large extent | 3.38 ± 1.28 | 3.44 ± 1.12 |
| *Possibilities for development* | Do you have the possibility of learning new things through completing tasks on MTurk/Prolific? | To a very small extent - To a very large extent | 3.42 ± 1.17 | 3.59 ± 1.10 |
| *Meaning of work** | Do you feel that completing tasks on MTurk/Prolific is meaningful? | To a very small extent - To a very large extent | 3.78 ± 1.05 | 3.48 ± 1.18 |
| *Social support from requesters** | How often do you get help and support from MTurk/Prolific or task requesters, if needed? | Never - Always | 2.69 ± 1.30 | 3.02 ± 1.18 |
| *Social support from workers** | How often do you get help and support from other workers, if needed? | Never - Always | 1.96 ± 1.21 | 2.97 ± 1.19 |
| *Feedback at work** | How often do task requesters bonus/message you because how well you carry out your work? | Never - Always | 2.44 ± 0.96 | 3.30 ± 0.89 |
| *Sense of community** | Is there a good atmosphere between you and other workers (on either crowdsourcing platforms or other worker forums e.g. Reddit)? | To a very small extent - To a very large extent | 3.15 ± 1.02 | 3.44 ± 1.02 |

results compared to physical health (Figure 5.16). According to Shapiro-Wilk tests, we found that workers' emotional wellbeing scores do not come normal distributions ($p < 0.05$ for both platforms), while working energy/fatigues scores do ($p = 0.10$ and $p = 0.41$ for Prolific and MTurk respectively). Therefore, we applied the Mann-Whitney U test to test worker emotional wellbeing and the independent t-test to test worker energy/fatigue. Significant differences were found in both emotional wellbeing ($p = 0.001, CL = 0.40$) and energy/fatigue ($p = 0.014$, Cohen's $d = 0.30$). Therefore, while the MTurk workers are found healthier physically, the mental health scores with regard to both emotional wellbing and energy/fatigue are significantly lower than the Prolific workers.

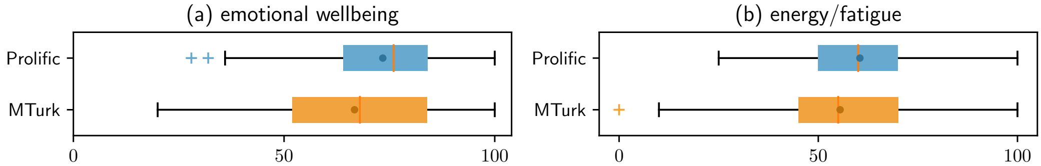


Figure 5.17: Health scores of workers' mental wellbeing and energy/fatigue (subsets of SF-36) across two crowdsourcing platforms (Prolific and MTurk).

To find factors that may potentially affect mental health, we conducted Spearman's correlation analysis between the psychosocial condition (in 10 dimensions) and mental health scores (of mental wellbeing and fatigue/energy). The results of correlation analysis (the correlation coefficients $\rho$ and $p$-values are reported in Table 5.9. Results showed that $\rho$

values are mostly less than 0.3, meaning relatively weak correlations in general. However, among all the relations, we found that the job content, in terms of the meaning of work ($\rho = 0.34, p < 0.001$) and the possibility of development (representing learning in the context of crowdsourcing, $\rho = 0.23, p < 0.001$), may affect working energy/fatigue; while the type of production, in terms of emotional demands ($\rho = 0.32, p < 0.001$) and work pace ($\rho = 0.21, p < 0.001$), may affect workers' emotional wellbeing. Both crowdsourcing platforms and task requesters should consider these factors in the future task design, by for instance, emphasizing the meaning of crowdsourced work in task batches and involving more learning elements [29, 117].

Table 5.9: The Spearman's rank correlation coefficients ($\rho$) of relations between the self-reported health scores and the psychosocial conditions. Asterisk (*) represents significance.

| | **General health** | | **Emotional wellbeing** | | **Energy/Fatigue** | |
|---|---|---|---|---|---|---|
| | $\rho$ | $p$-value | $\rho$ | $p$-value | $\rho$ | $p$-value |
| ***Quantitative demands*** | 0.059 | 0.325 | 0.086 | 0.149 | 0.083 | 0.164 |
| ***Work pace*** | -0.074 | 0.214 | 0.211 | <0.001* | 0.168 | 0.005* |
| ***Emotional demands*** | -0.074 | 0.217 | 0.322 | <0.001* | 0.144 | 0.015* |
| ***Influence at work*** | 0.097 | 0.104 | 0.086 | 0.151 | 0.195 | 0.001* |
| ***Possibilities of development*** | 0.125 | 0.036* | 0.095 | 0.112 | 0.230 | <0.001* |
| ***Meaning of work*** | 0.201 | 0.001* | 0.182 | 0.002* | 0.342 | <0.001* |
| ***Social support from requesters*** | 0.017 | 0.775 | -0.037 | 0.541 | 0.120 | 0.044* |
| ***Social support from workers*** | 0.100 | 0.094 | -0.110 | 0.065 | -0.041 | 0.488 |
| ***Feedback at work*** | 0.229 | <0.001* | 0.029 | 0.627 | 0.161 | 0.007* |
| ***Sense of community*** | 0.197 | 0.001* | 0.109 | 0.067 | 0.176 | 0.003* |

***Summary of mental health analysis:***

1. MTurk workers reported worse mental health status compared to Prolific workers, whereas they reported better physical health status.

2. Crowdsourcing tasks on MTurk were more emotional demanding, and required workers to work faster. Prolific workers believed crowd work was more meaningful. However, the MTurk workers received more social supports/feedback from the platform/requesters/other workers.

3. The job content in terms of the meaning of work and the possibility of learning relates to workers' fatigue and energy, and task demands and the work pace may affect workers' emotional wellbeing.

### Worker Needs

Understanding the needs of workers is necessary for further improving worker health. In the fourth part of the survey, we explicitly asked workers what did they need concerning interventions to improve their health while completing crowdsourcing tasks.

In terms of workers' needs about physical health, aligned with self-reported physical discomfort, a very large proportion of the workers on both platforms would like to receive physical exercises and instructions for their necks, shoulders, and backs, as shown in Figure 5.18 (a). In terms of needs about mental health, as we can see from Figure 5.18 (b), Prolific workers and MTurk workers shared similar preferences. The aspect of mental health and wellbeing was the most aspired by workers. For all the other aspects, there were a considerable number of workers would like to see them in the interventions.

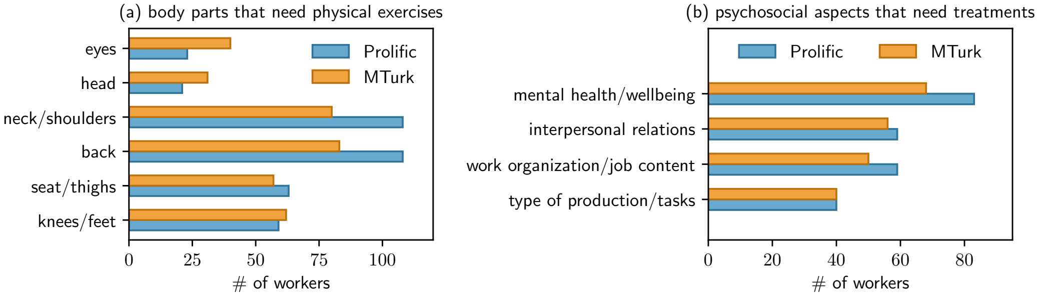


Figure 5.18: Worker needs in terms of (a) physical health and (b) mental health respectively across two crowdsourcing platforms (Prolific and MTurk).

In this study, we did not look into how health-related exercises or treatments could be realized and implemented on crowdsourcing platforms. We were more interested in to what extent workers would be happy to receive the interventions. As displayed in Figure 5.19, 82 workers (61%) on MTurk and 81 workers (55%) on Prolific, respectively, reported that they would be happy to use a tool that can sent them interventions to a (very) large extent. Particularly, Prolific workers would like to receive simple breaks and exercises for physical discomfort, while MTurk workers needed physical exercises the most. Only 7 workers from Prolific and 11 workers from MTurk did not want to receive interventions at all.

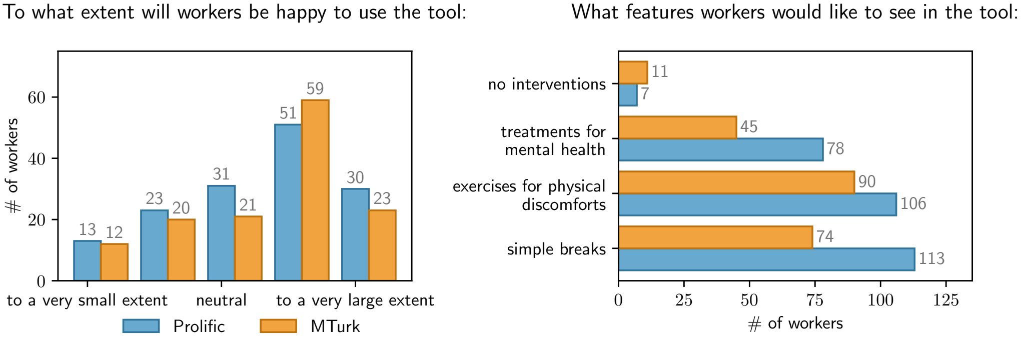


Figure 5.19: Worker needs with regard to a health-related tool that can provide interventions (breaks/exercises/treatments) across two crowdsourcing platforms (Prolific and MTurk).

The preferences of the workers with regard to different features of the interventions are shown in Figure 5.20. We explicitly asked workers about their opinions and preferences in terms of the working mode (push/pull interventions), payment mode, intervention timing, intervention length, preferred frequency, and who should be responsible for developing a tool to provide interventions to improve their health.

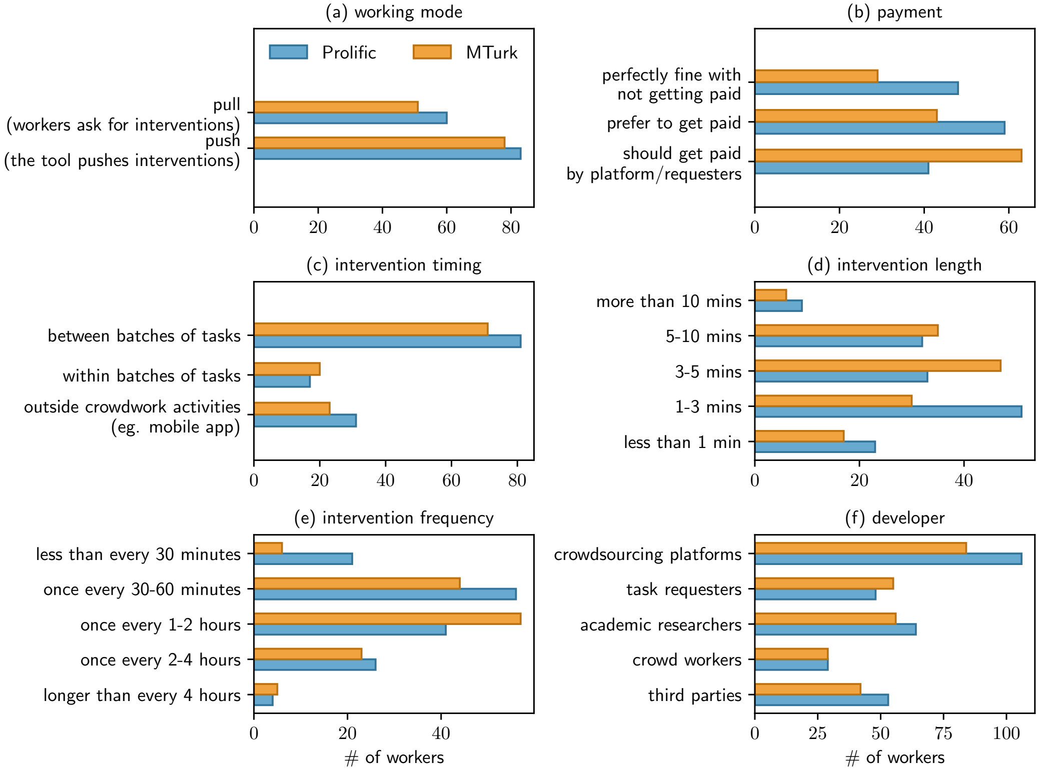


Figure 5.20: The features of the tool in terms of (a) working modes, (b) payment, (c) intervention timing, (d) intervention length, (e) intervention frequency, and (f) developer that workers preferred across two crowdsourcing platforms (Prolific and MTurk).

Working mode represents the way that the worker prefer to receive the interventions, where "pull" means workers asks for interventions (breaks/exercises/treatments) when they want them and "push" means the tool actively pushes interventions to workers. Results revealed that, as shown in Figure 5.20 (a), workers on both platforms preferred the "push" mode meaning they needed a tool reminding them of taking breaks/exercises/treatments. In terms of payment mode (Figure 5.20 (b)), a large proportion (63%) of workers on both platforms reported that they were okay with not getting paid (of these workers, 102 preferred to get paid and 77 were perfectly fine with not getting paid). As for intervention timing (Figure 5.20 (c)), representing when/where workers would like to receive interventions, the majority of the workers (152 across two platforms) wanted to take breaks/exercises/treatments between batches of tasks. As shown in Figure 5.20 (d), the length of intervention minutes should not be neither too long (> 10 minutes) nor too short (< 1 minute). As we can see Figure 5.20 (e), similarly, according to workers' answers, the intervention should not be sent neither too frequently (every < 0.5 hours) nor too infrequently (every > 4 hours). Finally, most workers on both platform "agreed" that the crowdsourcing platforms should be responsible for developing such tool/functions for sending interventions to improve their health. This finding corresponds to the result of intervention timing shown in Figure 5.20

(c), where interventions between batches of tasks implies that interventions should be at the platform level.

***Summary of worker needs:***

1. Workers on both platforms had similar needs with regard to interventions (breaks/exercises/treat during task execution) for improving their health.
2. Most workers would be happy to receive interventions, and especially would like to receive breaks and physical exercises (for necks/shoulders/backs), which were not longer than 10 minutes, between batches of tasks, every 0.5-2 hours.
3. Workers preferred that the interventions could be actively pushed to them.
4. Most workers were fine with not getting paid while taking breaks/exercises/treatments.
5. Most workers believed that the crowdsourcing platform should take the greatest responsibility of providing interventions.

### Discussion

Workers from MTurk and Profilic reported similar health-related issues. In terms of physical health, crowd workers from both platforms felt less comfortable in their necks, shoulders and backs. The physical discomfort was found to be related to the working ergonomics, meaning an unhealthy working posture can very likely cause physical discomfort. Furthermore, MTurk workers in our study reported a relatively better physical health status but relatively worse mental health status compared to Prolific workers. In the following section, we explore workers' reasons, comments, and suggestions to get a better understanding of worker health on crowdsourcing platforms.

### Insights from Workers' Feedback

***Health measures.*** In the first part of the survey, we explicitly asked workers whether they took some measures to keep healthy, before asking questions about their physical and mental health. We manually inspected all the comments of 283 valid submissions, and carried out a thematic analysis of the open-ended responses [207]. 61 workers reported they did not take any measures. Among responses from the other workers, the most popular measures reported included – drinking water regularly, eating healthy, taking breaks, and exercising frequently. Selected comments listed below, illustrate these insights from the workers.

> *Work out, drink water, go on daily walks, vitamin-pills, take as much vitamin D as possible (sun).* — Prolific worker, Switzerland, Female, Age 18-25.

> *I try to move around a few times an hour to relieve pressure from sitting and I also try to stay hydrated and eat nutritious snacks.* — MTurk worker, United States, Male, Age 36-45.

Two workers reported that they were already using ergonomic devices:

> *All of my devices are ergonomic, including my chair. Also I try to eat healthy and at least exercise a bit when I have been on the computer for too long.* — Prolific worker, Mexico, Male, Age 18-25.

> *I don't take long times of work without pause, I stretch regularly, try to eat healthy and I bought a chair with good back support to correct posture.* — Prolific worker, Chile, Male, Age 26-35.

In general, most workers did take some simple measures to make themselves healthy (around 80%). A few workers had already paid attention to working ergonomics and tried to ensure usage of ergonomic devices. However, many workers reported never taking measures to safeguard their health (around 20% workers), and one worker reported serious health-related problems (neuropathy and lupus).

***Needs for interventions.*** As can be seen in the results, shown in Figure 5.19 (a), most workers would be happy to receive interventions (163 out of 283). When we asked for reasons, most workers gave general opinions that they thought interventions were helpful. However, 13 workers reported somatic problems, and 11 workers reported not feeling well mentally.

> *Sometimes my back hurts but it's because my posture is bad on the chair, so sometimes I try to fix it but its hard.* — Prolific worker, Portugal, Male, Age 18-25.

> *I think that crowdsource working like on MTurk can be very sedentary, lonely and sometimes depressing, so some tools to help improve overall health would be good!* — MTurk worker, United States, Male, Age 36-45.

> *To many people (or task givers) treat us workers like we are not really important most of the time. Short time to complete tasks, and very low pay grade. All the support crowd source workers can get really helps us out a lot.* — MTurk worker, United States, Male, Age 36-45.

We are also interested in why there are still a considerable number of workers (68 out of 283)who preferred to receive interventions to small extent. The main reason is that many workers do not work very long on the crowdsourcing platform (especially Prolific) which makes the intervention not really necessary.

> *This is just something I do to fill time and for fun. I don't need to be interrupted. I just want to do the task and then get on with real work. It would frustrate me to be told to have breaks. I do a survey, have a snack then do my real job.* — Prolific worker, United Kingdom, Male, Age 36-45.

> *I don't spend a lot of time doing this so, I'd probably don't find it very useful but I think that tool would very important for someone who spends much more time working on crowdsourcing platforms, especially because in this time we face a lot of mental health problems and people still see this problems as a tabu even when mental health kills millions of people every year.* — Prolific worker, Portugal, Male, Age 18-25.

Unfortunately, we also observed 7 workers reported a reason that is very worrying – since some workers really need to earn money to support their livelihood, they were reluctant to the idea of spending time to improve their own health.

> *Lack of time. HITs are long and pay little. In order to earn enough money, I don't have time for breaks or exercises.* — MTurk worker, United States, Male, Age 36-45.

> *I don't think it's a bad idea at all, but I'm not sure if it will be attractive to crowd workers, since it's more optimal to spend all the time they have doing tasks.* — Prolific worker, Spain, Male, Age 18-25.

***Workers' preferred content of interventions.*** As shown in Figure 5.19 (b), most workers (187 out of 283) would like to receive breaks, particularly for Prolific workers, since them did not work long and did not have many serious health-related issues.

> *As I explained above the simple breaks between the study is enough to get my mind relaxed and continue to do the task and give properly formulated answers.* — Prolific worker, Portugal, Male, Age 26-35.

For some workers, physical exercises and mental treatments are also of a high priority (79 workers).

> *Physical discomforts are the highest priority for me. They would be the most concerning to alleviate.* — MTurk worker, United States, Male, Age 36-45.

> *Mental health should be a priority especially in these troubled times. And exercise can help with that too.* — Profilic worker, Chile, Male, Age 26-35.

***Workers' preferred working mode.*** We asked workers to select their preferred working mode of getting interventions. A push mode means that the interventions are actively pushed to workers, while a pull mode represents that workers ask for interventions when they want to. The majority of the workers preferred the push mode (161 out of 283), since they believed the intervention should work as an alert to remind them of taking some breaks or exercises.

> *I often download some tools for productivity and they usually just suggest to stop scrolling on the phone, so I just ignore them. If the tool actively pushes you to do that you can't ignore it, and if you downloaded that app then you are committed enough to let it guide you.* — Prolific worker, Portugal, Male, Age 18-25.

There were 111 workers (39%) preferred a pull mode. The main reason that workers would like to use a pull mode was that they wanted to decide when to take interventions by themselves.

> *I would not want to be interrupted with a notification to take a break. If I need one I'll take one in between surveys.* — Prolific worker, United States, Female, Age 46-55.

In addition, some workers (31 out of 283) would like the working mode to be more flexible and customized.

> *I might forget to ask for a break, that is the whole point, a reminder would be helpful. But sometimes I can't stop working without notice, so it would be good to be able to make my own schedule too.* — MTurk worker, United States, Female, Age 56-65.

***Workers' preferred payment mode.*** As reported in Figure 5.20 (b), we found a large proportion of workers were fine with not getting paid for restorative activities (179 out of 283). The main reason is that taking breaks is not the part of work, but they would prefer to get paid since it could be very strong incentive.

> *I would not explicitly demand to be paid, but money is always a good bonus.* — Prolific worker, United Kingdom, Male, Age 18-25.

> *It seems to me that taking breaks while doing crowdsourcing tasks should be my responsibility, as they usually aren't similar to having a regular job timewise or otherwise.* — Prolific worker, Poland, Female, Age 26-35.

> *I would be nice to get paid since the exercise is taking up my normal work time when I would be getting paid. However, the exercises would directly benefit me, so I would be open to doing them without getting paid.* — MTurk worker, United States, Female, Age 46-55.

Furthermore, as we can also see from Figure 5.20 (b), MTurk workers tend to believe that they should get paid (47%, 63 out of 135 on MTurk). The possible reason might be that low payment issues are quite common on MTurk.

> *We push ourselves to the point of having issues so that we can make as much money as possible. Many of the times that we neglect ourselves and breaks is because we don't want to miss out on money. The reason we are having physical and mental issues is because of the work and how hard we have to work for little pay.* — MTurk worker, United States, Female, Age 36-45.

***Free comments and suggestions.*** The most common signal that we received from comments is that workers found the idea of improving worker health useful and important. However, many workers also emphasized that low pay is the major issue that they are encountering in crowdsourcing marketplaces.

> *I think it's a fantastic idea to try and promote wellness, but it really is vital to emphasise how low pay will affect these things. If you make them mandatory, they'll irritate people and result in lower pay; if you don't pay for them, people may feel they have to skip them to keep working, which would make it an empty gesture at best.* — Prolific worker, United Kingdom, Male, Age 26-35.

Therefore, once again, we want the crowdsourcing platforms and task requesters to pay more attention to the wellness of crowd workers, especially when we found the following comment:

> *The only thing that needs improvement on mTurk is helping us - workers negotiate our rejections with Requesters. I got so many unfair rejections, and NOBODY from Amazon is helping me and Requesters simply don't answer emails...* — MTurk worker, United States, Male, Age 36-45.

Furthermore, it is worth mentioning that we found that some workers were fond of completing online survey with conversational crowdsourcing (although we did not ask for their opinions about it), since it made them felt more engaged and less bored, which is aligned with previous research.

> *I really enjoyed the chatbot format of the survey, it makes it feel more personal and less tedious than other formats.* — Prolific worker, Mexico, Male, Age 18-25.

> *This type of survey (in form of 'texts' and not several pages of questions) made me more invested and less tired mentally, and thank you for caring about us, mturk workers, I wish you the best day.* — MTurk worker, Brazil, Female, Age 18-25.

### General Health Status of Workers

Results pertaining to the health analysis revealed that workers across both platforms, share a common pattern of health status. First of all, we found that a large proportion of the workers worked alone, meaning that they either did not have any co-workers (or were not aware of crowd worker communities) or they did not share workplaces with others. Furthermore, workers on both platforms reported that they felt less comfortable in their necks, shoulders, and backs, which relate to the common somatic pain caused by long-time office work [177]. Our findings suggest that an unhealthy working posture could potentially be the main cause of body discomforts experienced by crowd workers. As for mental health, we found that the job content in terms of the meaning of work and the possibility of learning could affect workers' fatigue and energy, while the task demands and the work pace could affect workers' emotional well-being.

Differences in the quality of the working environment could be explained by the more prominent role that crowd work plays for workers in the two platforms. MTurk workers reported a longer working time on average, and more crowdsourcing experience in comparison to Prolific workers. On average, MTurk workers also reported having relatively better working environments with respect to lighting, temperature, humidity, noise, devices, control, posture, and so forth.

On Prolific, workers usually do not spend much time on completing microtasks, while on MTurk, workers spend a much longer time completing tasks that are in general more demanding. Due to the well-documented platform dynamics on MTurk, workers usually tend to complete tasks at a higher pace to optimize their earnings [57, 88, 192]. Such differences could help explaining why MTurk workers show a worse status with regard to mental wellbeing and working energy, compared to Prolific workers. We also found that crowd workers from MTurk have a better sense of community and acquire more social supports from outside the platform itself. This is absolutely a good sign, but we must admit that many turker communities and unities were found exactly because of the poor working situation.

### Design Implications

***For crowdsourcing platforms.*** The results of this study clearly indicated that most workers would be happy to see a healthcare function integrated into their work routines, or embedded on the crowdsourcing platform. Workers on both platforms elicited similar needs with regard to receiving interventions, including simple breaks, physical exercises, or mental treatments, for improving their health. Our survey has shown that it would be appropriate to design and provide health interventions actively to workers (lasting no longer than 10 minutes), every 0.5-2 hours, between batches of tasks. Nevertheless, the types of interventions, their duration, content, and frequency of the interventions should be customizable and personalized to worker preferences. Workers who do not prefer to receive such interventions should have an easy and accessible way to opt-out of them.

Currently, crowdsourcing platforms act as an intermediary agent that only introduces jobs and tasks to workers, where the work and jobs are not properly supervised nor legally protected. As more and more people turn towards crowdsourcing marketplaces (and the broader spectrum of online work) to support their livelihood, task requesters and crowdsourcing platforms alike, should gradually take the responsibility of safeguarding the health and wellbeing of crowd workers. We envision a future of crowd work in which crowdsourcing platforms design provisions to sustain a healthy workforce. Apart from crowning workers with qualifications, and virtual badges to reward their long-term and high-quality work, platforms can consider rewarding workers with ergonomic devices to support their continued work or provide them with necessary health interventions. After all, few factors may contribute more to the sustainable growth and prosperity of a paid crowdsourcing platform than fostering a healthy relationship with crowd workers and ensuring their wellbeing. However, several important question need to be addressed before such a reality can be realized. How can health interventions be introduced and packaged between or within HITs? To what extent would such interventions serve as an effective means to improve worker health and wellbeing as a result of crowd work?

***For task requesters.*** According to the results of mental health questions, the task content also plays an important role in worker health. The meaning of work and the possibility of learning can significantly affect workers' mental health. A simple starting point for task requesters is to emphasize the meaning of the task, rather than immediately giving them instructions and letting them work like robots, as suggested by Chandler et al. [29]. Furthermore, also as suggested by previous work [117], task requesters should consider involving more learning elements in crowd work, which has been proved to be effective in terms of better engaging workers and improving their performance [52]. The task demands and the work pace were also found to be related to workers' emotional wellbeing. Heavy task demands and a fast work pace could be caused by worker themselves since many of them would like to earn more money using limited time. However, for task requesters, it is not necessary to further give pressure on crowd work. A relatively loose working environment and time limit could have positive effects on workers' mental wellbeing. Furthermore, content moderation on social media heavily relies on crowd work, meaning workers (i.e. content moderators) might be exposed to harmful content for a long time. This kind of emotional demands could affect workers' emotional wellbeing according to prior studies [204], which is also in-line with our results. Task requesters should also take this into account during the task design phase [44].

Task requesters can begin by shouldering some of the responsibility to ensure worker wellbeing. We envision that requesters can provide crowd workers with small health interventions, designed as tasks to be completed and packaged together with HITs. Paying crowd workers to consume such health interventions would result in increasing the costs for the task requesters by a relatively small fraction. However, in return task requesters can reap the benefits of having a healthy and sustainable workforce to rely on and mutually flourish.

### Limitations and Future Work

In this study, we tried recruiting 300 workers from two platforms (MTurk and Prolific). After excluding unreliable submissions, we had 283 workers for analysis. We acknowledge that the recruited participants could be only partially representative of the overall population of the selected crowdsourcing marketplaces. Future work could recruit more participants and involve more crowdsourcing platforms (such as Appen[39] and Toloka[40]), or consider performing studies on online freelancing marketplaces, to make broader implications with regard to the entire online gig economy.

Furthermore, future work could focus on systematic and detailed research of a specific health-related aspect, such as working ergonomics, somatic discomforts, psychosocial working environment and so on. For instance, to comprehensively assess the psychosocial working environment, at least 44 questions are needed in a short-version COPSOQ questionnaire. In this work, since it is the first step towards understanding worker health, we did not try to go deeper into each aspect. Using verified surveys to systematically assess worker health could be a promising research direction in the imminent future. It would be meaningful and also valuable to conduct profound studies concerning worker health among different groups of workers. The health status of crowd workers could be compared according to their genders, ages, countries, working experiences, etc. For example, it would be interesting to explore whether crowd work experience can help workers create healthier working environments for themselves.

[39] `https://appen.com/`

[40] `https://toloka.yandex.com/`

## 5.3 Chapter Summary

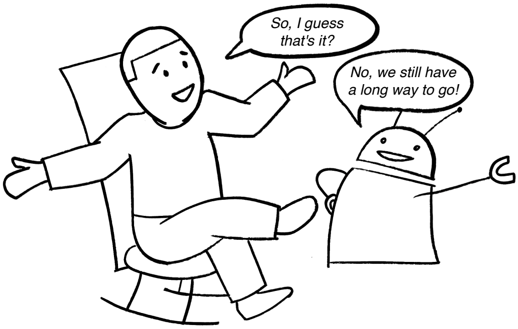
This chapter presents two applications (other than microtasking) of conversational crowdsourcing. The first application of conversational crowdsourcing (Section 5.1) presents a first exploration of how human memorability can be improved in information retrieval. To this end, we proposed novel search interfaces and quantified long-term memorability. We designed user interfaces with text-based conversational agents and note-taking affordances for informational search. In the second application of conversational crowdsourcing (Section 5.2), we investigated worker health in crowdsourcing marketplaces by performing a conversational survey study on two crowdsourcing platforms – Prolific and MTurk respectively. We acquired data and performed analysis about workers' background, physical health status, and mental health status. Furthermore, we explicitly asked workers about their needs, and proposed suggestions to crowdsourcing platforms and task requesters for improving worker health.

The completion of the two studies shows the feasibility of applying conversational crowdsourcing in human-centered experiments. In the first application about improving user memorability in information retrieval, we show that conversational interfaces have the potential to benefit long-term memorability, which also indicates that conversational crowdsourcing can be effectively applied in the domain of information retrieval, to assist researchers in carrying out user studies relating to conversational interfaces. In terms of the second application about worker health, we found evidence suggesting that an unhealthy working posture is a likely cause for their physical discomfort, and we highlight that worker mental health can be affected by task content, work pace, and task emotional demands. Our findings suggest that for a healthy future of crowd workers, crowdsourcing platforms need to take major responsibility, together with task requesters and academic researchers, in providing healthcare interventions during crowd work to improve worker health. Workers' comments further reflect that conversational crowdsourcing can be applied in common human-centered experiments to achieve better worker satisfaction and engagement.

# Chapter 6

# Conclusions

TASK INSTRUCTIONS
Chapter 1
Introduction
Chapter 2
Designing Conversational Crowdsourcing
Chapter 3
Engagement and Satisfaction
Chapter 4
The Roles of Mood and Self-Identification
Chapter 5
Conversational Crowdsourcing Applications



Crowdsourcing has become increasingly important for effectively collecting human input from anonymous online workers, to build machine learning datasets, evaluate AI systems, and carry out online user studies. As crowdsourcing advances, lowering the participation barrier, increasing worker satisfaction, and improving worker engagement have been identified as major challenges to tackle. To this end, we introduce a novel interaction paradigm – conversational crowdsourcing – to effectively improve worker satisfaction and engagement, by proposing a set of methods combining conversational interfaces, theories of conversational styles, worker emotions, and self-identification with avatars inspired by games research.

## 6.1 Summary of Contributions

We have tackled the problem from four main angles: 1) *conversational crowdsourcing design*, 2) *improving worker engagement and satisfaction*, 3) *analyzing the roles of worker mood and self-identification*, and 4) *applying conversational crowdsourcing*. Each of the four angles is addressed by an individual chapter of this thesis.

### Designing Conversational Crowdsourcing

To address **RQ2.1** and **RQ2.2**, we designed conversational crowdsourcing by proposing novel workflows for conversational agents to support task execution. We carried out an experiment to thoroughly analyze the role that conversational crowdsourcing could play across five task types.

We show that conversational crowdsourcing can lead to similar task execution times and output qualities, in comparison to the traditional web interfaces. The findings suggest that conversational crowdsourcing is an effective alternative to the prevailing standard web-based crowd work. As our initial experiments required extra effort redirecting workers from the crowdsourcing platform to Telegram, to further lower the barrier of participation in crowdsourcing, we designed a tool assisting task requesters in quickly deploying conversational crowdsourcing tasks on web-based crowdsourcing platforms, named TickTalkTurk. This tool reduces the overheads of designing and implementing conversational interfaces for crowdsourcing.

### Improving Worker Engagement and Satisfaction

To address **RQ3.1**, **RQ3.2**, **RQ3.3** and **RQ3.4**, we carried out two experiments to study the effect of conversational crowdsourcing in improving engagement and satisfaction, and to explore the roles that conversational styles could play.

We show that, in general, conversational crowdsourcing can improve worker retention and perceived engagement. Particularly, a High-Involvement style can better engage workers. Furthermore, we proposed a conversational style estimation method, which can be applied in general conversational systems on common social network channels for analyzing and understanding the personality, emotion, and subjectivity of online users. We also show that workers tend to exhibit different conversational styles due to the effect of task complexity. We provide insightful implications for fostering a healthy relationship between workers and requesters in microtask marketplaces.

### Analyzing the Roles of Mood and Self-identification

To address **RQ4.1**, **RQ4.2**, **RQ4.3** and **RQ4.4**, we investigated the role of worker mood in the context of conversational microtask crowdsourcing, and the effect of self-identification with worker avatars in conversational crowdsourcing for information retrieval.

Our results clearly indicate that conversational crowdsourcing can improve worker retention irrespective of worker moods. We found that workers in pleasant moods reported better engagement scores, calling for dynamic task assignment strategies to induce pleasant moods prior to task execution. Interestingly, the findings suggest that conversational agents with a High-Considerateness style exhibit the potential to improve engagement of unpleasant workers, while a High-Involvement style exhibits a potential to further engage pleasant workers. In terms of self-identification with worker avatars, we show that using worker avatars with conversational crowdsourcing in complex tasks combining learning elements can lead to a reduction in the perceived cognitive load. Our results also reveal the occurrence of similarity identification and wishful identification in conversational crowdsourcing. The findings have important implications in future task design.

### Applying Conversational Crowdsourcing

To address **RQ5.1**, **RQ5.2**, **RQ5.3** and **RQ5.4**, we first applied conversational crowdsourcing in the realm of information retrieval by carrying out a user study to investigate the effectiveness of conversational interfaces in improving human memorability, and then applied conversational crowdsourcing by carrying out a survey study to understand worker health on popular crowdsourcing platforms.

Our findings reveal that users employing conversational interfaces in informational search sessions exhibit a different search behavior compared to traditional web search. We find that a conversational interface has the potential to help users better retain information consumed. These findings suggest that conversational interfaces can be a promising tool towards memorable information retrieval. The experiment has shown the feasibility of applying conversational crowdsourcing in information retrieval for conducting user studies. Inn terms of worker health, we found that Mechanical Turk workers reported better physical health, while Prolific workers reported better mental health. Furthermore, the analysis reveals that: physical discomfort is related to the working ergonomics; energy levels can be affected by task content; and the mental wellbeing can be affected by work pace and task demands. Our survey emphasizes the importance of designing and providing health interventions on crowdsourcing platforms in the future. Furthermore, it is worth mentioning that workers reported that completing online survey with a conversational interface made them feel more engaged and less bored, which is in-line with our previous findings.

## 6.2 The Future of Conversational Crowdsourcing

This thesis shows the potential of conversational crowdsourcing for the better future of crowd work. We contribute novel techniques for lowering barriers for participation, increasing worker satisfaction, and improving worker engagement. While we consider our results promising, there is still room for improvement and further exploration. In this section, we identify several directions for further research.

### Worker-Centred Task Design

Although we have learnt from the thesis that employing a variety of UI elements such as conversational interfaces and avatar customization can improve worker satisfaction and engagement, we still highlight the importance of task-specific interaction design. We found that the task content could play an important role in worker experience and even in shaping their mental health. Current task design immediately gives workers instructions and let them work like robots in assembly lines [29]. Future task design could consider using conversational crowdsourcing to instruct, guide, or train crowd workers with better perceived engagement and satisfaction. Our experiments in Section 3.2, Section 4.2, and Section 5.1 suggest that conversational interfaces with learning elements could result in better worker experience and memorability. Since the importance of involving learning elements in crowd work has also been identified in previous work [117] and the combination of conversational interfaces and learning shows positive effects in this thesis, in the imminent future, introducing more possibilities of learning and personal development in conversational crowdsourcing should be taken into account in task design. For instance, this could be realized by crowd workers getting qualifications through learning to access more tasks and earn more rewards, or by online learners gaining more learning resources through completing crowd work.

### Lowering Participation Barriers in Conversational Crowdsourcing

In this thesis across all the chapters, we only focus on text-based conversation. Text-based conversation is the most common way being used in our daily messaging applications. Nevertheless, text-based conversation ignores paralinguistic features like pitch and voice, non-linguistic features such as smile, laughter and gestures, and it also restricts the participation for crowd workers with disabilities. Furthermore, a recent work has also pointed out the importance of conversation backchannels in human-computer interaction [106]. Meanwhile, we have noticed a gradual rise in the use of audio- or video-based conversation interfaces. To further investigate the potential of conversational crowdsourcing and improve its accessibility, future work could explore various means to complete crowdsourcing tasks with conversational agents. For example, in the imminent future, we can investigate the feasibility of using voice-based conversational agents to assist crowd workers in completing crowdsourcing tasks, and explore the suitability of analyzing conversational styles using audio data.

### Fostering Intrinsic Motivation

Worker motivation has been identified as an important factor to achieve better worker satisfaction [117]. Across all the experiments carried out in the thesis, workers earned monetary rewards. Since our goal through conversational crowdsourcing is to improve worker satisfaction and engagement and to help workers overcome fatigue or boredom and reduce task abandonment, the future of conversational crowdsourcing could focus on workers' intrinsic motivations. Intrinsic motivation has already been playing an important role in gaming systems. In gaming systems, players are mainly motivated by enjoyment and interest when they are playing games, whereas in crowd work, workers are mostly motivated by monetary incentives when they are completing microtasks. Nevertheless, accroding to the results of Section 4.2, using avatar customization in conversational crowdsourcing does not significantly impact workers' intrinsic motivations according to the statistical tests. Based on our findings in the paid microtask crowdsourcing setup, future work could explore other means to foster intrinsic motivation by researching stronger motivational task design by leveraging the advantages of conversation. For instance, a promising research direction could be applying conversational interfaces in voluntary crowdsourcing campaigns with the game with a purpose (GWAP) design.

### Improving Worker Health

Improving worker health is a crucial step towards the future of crowd work. The study carried out in Section 5.2 did not try to go deeper into each aspect of worker health (working ergonomics, somatic discomforts, psychosocial working environment, etc.). Future work could use standardized surveys to systematically assess worker health. The health status of crowd workers could be analyzed according to their genders, ages, countries, working experiences, etc., to provide personalized support. Furthermore, we did not design a tool to provide health interventions on crowdsourcing platforms, since this is the first step of understanding worker health in crowd work marketplaces. Based on our findings about the correlations between conversational crowdsourcing and workers' emotions (such as satisfaction, engagement, and moods), our next step is to provide effective interventions to improve worker health. For instance, such health-related interventions could be designed in a conversational, customizable, and personalized way, which provide crowd workers with breaks, physical exercises, and mental treatments. The interventions could be published as large batches of tasks using conversational crowdsourcing in crowd work marketplaces.

# Bibliography

[1] Tahir Abbas, Vassilis-Javed Khan, Ujwal Gadiraju, and Panos Markopoulos. Trainbot: A Conversational Interface to Train Crowd Workers for Delivering On-Demand Therapy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 3–12, 2020.

[2] Ahmed Abbasi, Hsinchun Chen, and Arab Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):1–34, 2008.

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

[4] Richard C Atkinson and Richard M Shiffrin. Human memory: A proposed system and its control processes. *Psychology of learning and motivation*, 2:89–195, 1968.

[5] Sandeep Avula. Searchbots: Using Chatbots in Collaborative Information-seeking Tasks. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1375, New York, NY, USA, 2017. ACM.

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

[7] Alan D Baddeley, Neil Thomson, and Mary Buchanan. Word length and the structure of short-term memory. *Journal of verbal learning and verbal behavior*, 14(6):575–589, 1975.

[8] Seyed Ali Bahrainian and Fabio Crestani. Augmentation of Human Memory: Anticipating Topics That Continue in the Next Meeting. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, pages 150–159, New York, NY, USA, 2018. ACM.

[9] Christopher Beedie, Peter Terry, and Andrew Lane. Distinctions between emotion and mood. *Cognition & Emotion*, 19(6):847–878, 2005.

[10] Tara S Behrend, David J Sharek, Adam W Meade, and Eric N Wiebe. The viability of crowdsourcing for survey research. *Behavior research methods*, 43(3):800, 2011.

[11] Alberto Betella and Paul F M J Verschure. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PloS one*, 11(2), 2016.

[12] John B Biggs. *Student Approaches to Learning and Studying. Research Monograph.* Australian Council for Educational Research, Hawthorn, 1987.

[13] Jeffrey P Bigham, Kristin Williams, Nila Banerjee, and John Zimmerman. Scopist: building a skill ladder into crowd transcription. In *Proceedings of the 14th Web for All Conference on The Future of Accessible Work*, pages 1–10, 2017.

[14] Max V Birk, Cheralyn Atkins, Jason T Bowey, and Regan L Mandryk. Fostering intrinsic motivation through avatar identification in digital games. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2982–2995, 2016.

[15] Max Valentin Birk and Regan Lee Mandryk. Improving the efficacy of cognitive training for digital mental health interventions through avatar customization: crowdsourced quasi-experimental study. *Journal of medical Internet research*, 21(1):e10133, 2019.

[16] Tristan Blanc-Brude and Dominique L Scapin. What Do People Recall About Their Documents?: Implications for Desktop Search Tools. In *Proceedings of the 12th International Conference on Intelligent User Interfaces*, IUI '07, pages 102–111, New York, NY, USA, 2007. ACM.

[17] Gordon H Bower and Stephen G Gilligan. Remembering information related to one's self. *Journal of research in personality*, 13(4):420–432, 1979.

[18] Anne Bowser, Derek Hansen, Yurong He, Carol Boston, Matthew Reid, Logan Gunnell, and Jennifer Preece. Using gamification to inspire new citizen science volunteers. In *Proceedings of the first international conference on gameful design, research, and applications*, pages 18–25, 2013.

[19] Alessandro Bozzon, Marco Brambilla, Stefano Ceri, and Andrea Mauri. Reactive crowdsourcing. In *Proceedings of the 22nd international conference on World Wide Web - WWW '13*, pages 153–164, Rio de Janeiro, Brazil, 2013. ACM.

[20] Luka Bradeško, Michael Witbrock, Janez Starc, Zala Herga, Marko Grobelnik, and Dunja Mladenić. Curious Cat–Mobile, Context-Aware Conversational Crowdsourcing Knowledge Acquisition. *ACM Transactions on Information Systems (TOIS)*, 35(4):1–46, 2017.

[21] Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[22] Erin Brady, Meredith Ringel Morris, and Jeffrey P Bigham. Gauging receptiveness to social microvolunteering. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1055–1064. ACM, 2015.

[23] Jonathan Bragg and Daniel S Weld. Sprout: Crowd-powered task design for crowdsourcing. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pages 165–176, 2018.

[24] Dung C Bui, Joel Myerson, and Sandra Hale. Note-taking with computers: Exploring alternative strategies for improved recall. *Journal of Educational Psychology*, 105(2):299, 2013.

[25] John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on Twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1301–1309. Association for Computational Linguistics, 2011.

[26] Jenna Butler, Mary Czerwinski, Shamsi Iqbal, Sonia Jaffe, Kate Nowak, Emily Peloquin, and Longqi Yang. Personal Productivity and Well-being–Chapter 2 of the 2021 New Future of Work Report. Technical report, Microsoft, 2021.

[27] John T Cacioppo, Louise C Hawkley, L Elizabeth Crawford, John M Ernst, Mary H Burleson, Ray B Kowalewski, William B Malarkey, Eve Van Cauter, and Gary G Berntson. Loneliness and health: Potential mechanisms. *Psychosomatic medicine*, 64(3):407–417, 2002.

[28] Andrea Ceroni, Nam Khanh Tran, Nattiya Kanhabua, and Claudia Niederée. Bridging temporal context gaps using time-aware re-contextualization. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 1127–1130. ACM, 2014.

[29] Dana Chandler and Adam Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, 90:123–133, 2013.

[30] Adrian Chen. Inside Facebook's Outsourced Anti-Porn and Gore Brigade Where Camel Toes are More Offensive Than Crushed Heads. *Gawker. Com*, 16, 2012.

[31] Adrian Chen. The laborers who keep dick pics and beheadings out of your Facebook feed. *Wired*, 23:14, 2014.

[32] Zoey Chen and Jonah Berger. When, why, and how controversy causes conversation. *Journal of Consumer Research*, 40(3):580–593, 2013.

[33] Justin Cheng, Jaime Teevan, and Michael S Bernstein. Measuring crowdsourcing effort with error-time curves. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1365–1374, 2015.

[34] Chun-Wei Chiang, Anna Kasunic, and Saiph Savage. Crowd coach: Peer coaching for crowd workers' skill growth. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–17, 2018.

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

[36] Susanne Colenberg, Tuuli Jylhä, and Monique Arkesteijn. The relationship between interior office space and employee health and well-being–a literature review. *Building Research & Information*, pages 1–15, 2020.

[37] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Others. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760, 2010.

[38] E Nigel Corlett and R P Bishop. A technique for assessing postural discomfort. *Ergonomics*, 19(2):175–182, 1976.

[39] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 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*, page 43. ACM, 2017.

[40] Nicole Crenshaw and Bonnie Nardi. What's in a name? Naming practices in online video games. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play*, pages 67–76, 2014.

[41] Mihaly Csikszentmihalyi and Mihaly Csikzentmihaly. *Flow: The psychology of optimal experience*, volume 1990. Harper & Row New York, 1990.

[42] Peng Dai, Jeffrey M Rzeszotarski, Praveen Paritosh, and Ed H Chi. 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*, pages 628–638. ACM, 2015.

[43] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1):1–40, 2018.

[44] Anubrata Das, Brandon Dang, and Matthew Lease. Fast, Accurate, and Healthier: Interactive Blurring Helps Moderators Reduce Exposure to Harmful Content. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 33–42, 2020.

[45] Vincenzo Della Mea, Eddy Maddalena, and Stefano Mizzaro. Mobile Crowdsourcing: Four Experiments on Platforms and Tasks. *Distrib. Parallel Databases*, 33(1):123–141, mar 2015.

[46] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[47] Tangjian Deng, Liang Zhao, Ling Feng, and Wenwei Xue. Information Re-finding by Context: A Brain Memory Inspired Approach. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1553–1558, New York, NY, USA, 2011. ACM.

[48] Pieter M A Desmet, Martijn H Vastenburg, and Natalia Romero. Mood measurement with Pick-A-Mood: review of current methods and design of a pictorial self-report scale. *Journal of Design Research*, 14(3):241–279, 2016.

[49] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. Demographics and dynamics of mechanical Turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 135–143, 2018.

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

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

[52] Mira Dontcheva, Robert R Morris, Joel R Brandt, and Elizabeth M Gerber. Combining crowdsourcing and learning to improve engagement and performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3379–3388, 2014.

[53] M v Eeuwen. Mobile conversational commerce: messenger chatbots as the next interface between businesses and consumers. Master's thesis, TU Twente, 2017.

[54] Carsten Eickhoff and Arjen P de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval*, 16(2):121–137, 2013.

[55] Carsten Eickhoff, Christopher G Harris, Arjen P de Vries, and Padmini Srinivasan. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 871–880, 2012.

[56] Gilles O Einstein, Joy Morris, and Susan Smith. Note-taking, individual differences, and memory for lecture information. *Journal of Educational psychology*, 77(5):522, 1985.

[57] Kinda El Maarry, Kristy Milland, and Wolf-Tilo Balke. A fair share of the work? The evolving ecosystem of crowd workers. In *Proceedings of the 10th acm conference on web science*, pages 145–152, 2018.

[58] Stanford University Ergonomics. Stanford's computer workstation ergonomics self-evaluation form. `https://ehs.stanford.edu/forms-tools/computer-workstation-ergonomics-evaluation`.

[59] Shaoyang Fan, Ujwal Gadiraju, Alessandro Checco, and Gianluca Demartini. CrowdCO-OP: Sharing Risks and Rewards in Crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–24, 2020.

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

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

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

[63] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e19, 2017.

[64] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[65] Asbjørn Følstad, Petter Bae Brandtzaeg, Tom Feltwell, Effie L-C. Law, Manfred Tscheligi, and Ewa A Luger. SIG: Chatbots for Social Good. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, pages SIG06:1—-SIG06:4, New York, NY, USA, 2018. ACM.

[66] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. Modus operandi of crowd workers: The invisible role of microtask work environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):49, 2017.

[67] Ujwal Gadiraju and Gianluca Demartini. Understanding Worker Moods and Reactions to Rejection in Crowdsourcing. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, HT '19, pages 211–220, New York, NY, USA, 2019. ACM.

[68] Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. Crowd anatomy beyond the good and bad: Behavioral traces for crowd worker modeling and pre-selection. *Computer Supported Cooperative Work (CSCW)*, 28(5):815–841, 2019.

[69] Ujwal Gadiraju and Stefan Dietze. Improving learning through achievement priming in crowdsourced information finding microtasks. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 105–114, 2017.

[70] Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(4):1–26, 2017.

[71] Ujwal Gadiraju, Neha Gupta, In Karin Hansson, Tanja Aitamurto, Thomas Ludwig, and Michael Muller. Dealing with sub-optimal crowd work: Implications of current quality control practices. *International Reports on Socio-Informatics (IRSI), Proceedings of the CHI*, pages 15–20, 2016.

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

[73] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1631–1640, 2015.

[74] Ujwal Gadiraju, Patrick Siehndel, Besnik Fetahu, and Ricardo Kawase. Breaking bad: Understanding behavior of crowd workers in categorization microtasks. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 33–38, 2015.

[75] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 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*, pages 5–14. ACM, 2017.

[76] Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. Analyzing knowledge gain of users in informational search sessions on the web. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, pages 2–11, 2018.

[77] Snehal Gaikwad, Durim Morina, Rohit Nistala, Megha Agarwal, Alison Cossette, Radhika Bhanu, Saiph Savage, Vishwajeet Narwal, Karan Rajpal, Jeff Regino, and Others. Daemo: A self-governed crowdsourcing marketplace. In *Adjunct proceedings of the 28th annual ACM symposium on user interface software & technology*, pages 101–102, 2015.

[78] Bertil Gardell. Psychosocial aspects of industrial production methods. Technical report, Department of psychology, University of Stockholm (Psykologiska Institutionen, Stockholms Universitet), 1979.

[79] A Ghoshal. Microsoft sued by employees who developed ptsd after reviewing disturbing content. The next web, https://thenextweb.com/news/microsoft-sued-by-employees-who-developed-ptsd-after-reviewing-disturbing-content, 2017.

[80] E Bruce Goldstein. *Cognitive psychology: Connecting mind, research and everyday experience*. Nelson Education, 2014.

[81] Amy L Gonzales, Jeffrey T Hancock, and James W Pennebaker. Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37(1):3–19, 2010.

[82] Etienne Grandjean and Wilhelm Hünting. Ergonomics of posture - review of various problems of standing and sitting posture. *Applied ergonomics*, 8(3):135–140, 1977.

[83] Mary L Gray and Siddharth Suri. *Ghost work: how to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books, 2019.

[84] Tobias Greitemeyer and Silvia Osswald. Effects of prosocial video games on prosocial behavior. *Journal of personality and social psychology*, 98(2):211, 2010.

[85] Nathan Hahn, Shamsi T Iqbal, and Jaime Teevan. Casual Microtasking: Embedding Microtasks in Facebook. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 19. ACM, 2019.

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

[87] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. The impact of task abandonment in crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[88] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. A data-driven analysis of workers' earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2018.

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

[90] Bob Heller, Mike Proctor, Dean Mah, Lisa Jewell, and Bill Cheung. Freudbot: An investigation of chatbot technology in distance education. In *EdMedia+ Innovate Learning*, pages 3913–3918. Association for the Advancement of Computing in Education (AACE), 2005.

[91] DANULA HETTIACHCHI, NIELS VAN BERKEL, VASSILIS KOSTAKOS, and JORGE GONCALVES. CrowdCog: A Cognitive Skill based System for Heterogeneous Task Assignment and Recommendation in Crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction*, 4, 2020.

[92] E Tory Higgins. Self-discrepancy: a theory relating self and affect. *Psychological review*, 94(3):319, 1987.

[93] Chien-Ju Ho and Jennifer Vaughan. Online task assignment in crowdsourcing markets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, 2012.

[94] Joses Ho, Tayfun Tumkaya, Sameer Aryal, Hyungwon Choi, and Adam Claridge-Chang. Moving beyond P values: data analysis with estimation graphics. *Nature methods*, 16(7):565–566, 2019.

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

[96] Cynthia Hoffner. Children's wishful identification and parasocial interaction with favorite television characters. *Journal of Broadcasting & Electronic Media*, 40(3):389–402, 1996.

[97] Cynthia Hoffner and Martha Buchanan. Young adults' wishful identification with television characters: The role of perceived similarity and character attributes. *Media psychology*, 7(4):325–351, 2005.

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

[99] Ting-Hao Kenneth Huang, Joseph Chee Chang, and Jeffrey P Bigham. 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*, page 295. ACM, 2018.

[100] Ting-Hao Kenneth Huang, Walter S Lasecki, Amos Azaria, and Jeffrey P Bigham. "Is There Anything Else I Can Help You With?" Challenges in Deploying an On-Demand Crowd-Powered Conversational Agent. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 2016.

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

[102] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[103] Margaret J Intons-Peterson and JoAnne Fournier. External and internal memory aids: When and how often do we use them? *Journal of Experimental Psychology: General*, 115(3):267, 1986.

[104] Lilly C Irani and M Six Silberman. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 611–620, 2013.

[105] Lilly C Irani and M Six Silberman. Stories We Tell About Labor: Turkopticon and the Trouble with" Design". In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4573–4586, 2016.

[106] Vidit Jain, Maitree Leekha, Rajiv Ratn Shah, and Jainendra Shukla. Exploring Semi-Supervised Learning for Predicting Listener Backchannels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[107] Kevin Jepson. Conversations and negotiated interaction in text and voice chat rooms. *Language Learning & Technology*, 9(3):79–98, 2005.

[108] V K. Chaithanya Manam, Dwarakanath Jampani, Mariam Zaim, Meng-Han Wu, and Alexander J. Quinn. TaskMate: A Mechanism to Improve the Quality of Instructions in Crowdsourcing. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1121–1130, 2019.

[109] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1941–1944, 2011.

[110] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information retrieval*, 16(2):138–178, 2013.

[111] David Kember. The intention to both memorise and understand: Another approach to learning? *Higher Education*, 31(3):341–354, 1996.

[112] Shashank Khanna, Aishwarya Ratan, James Davis, and William Thies. Evaluating and improving the usability of Mechanical Turk for low-income workers in India. In *Proceedings of the first ACM symposium on computing for development*, page 12. ACM, 2010.

[113] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 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, pages 86:1—-86:12, New York, NY, USA, 2019. ACM.

[114] Youjeong Kim and S Shyam Sundar. Visualizing ideal self vs. actual self through avatars: Impact on preventive health outcomes. *Computers in Human Behavior*, 28(4):1356–1364, 2012.

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

[116] Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 453–456, Florence, Italy, 2008. ACM.

[117] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1301–1318, 2013.

[118] Alex S Koch, Joseph P Forgas, and Diana Matovic. Can negative mood improve your conversation? Affective influences on conforming to Grice's communication norms. *European Journal of Social Psychology*, 43(5):326–334, 2013.

[119] Michiel Kompier. The psychosocial work environment and health - what do we know and where should we go? *Scandinavian journal of work, environment & health*, pages 1–4, 2002.

[120] Tage S Kristensen, Harald Hannerz, Annie Høgh, and Vilhelm Borg. The Copenhagen Psychosocial Questionnaire-a tool for the assessment and improvement of the psychosocial work environment. *Scandinavian journal of work, environment & health*, pages 438–449, 2005.

[121] Pavel Kucherbaev, Azad Abad, Stefano Tranquillini, Florian Daniel, Maurizio Marchese, and Fabio Casati. CrowdCafe-Mobile Crowdsourcing Platform. *arXiv preprint arXiv:1607.01752*, 2016.

[122] Pavel Kucherbaev, Alessandro Bozzon, and Geert Jan Houben. Human Aided Bot. *IEEE Internet Computing*, 2018.

[123] Katsumi Kumai, Masaki Matsubara, Yuhki Shiraishi, Daisuke Wakatsuki, Jianwei Zhang, Takeaki Shionome, Hiroyuki Kitagawa, and Atsuyuki Morishima. Skill-and-stress-aware assignment of crowd-worker groups to task streams. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 6, 2018.

[124] Abhishek Kumar, Kuldeep Yadav, Suhas Dev, Shailesh Vaya, and G Michael Youngblood. Wallah: Design and Evaluation of a Task-centric Mobile-based Crowdsourcing Platform. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, MOBIQUITOUS '14, pages 188–197, ICST, Brussels, Belgium, Belgium, 2014. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

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

[126] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y S Lau, and Others. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258, 2018.

[127] Walter S Lasecki and Jeffrey P Bigham. Online quality control for real-time crowd captioning. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, pages 143–150, 2012.

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

[129] Annabel Latham, Keeley Crockett, David McLean, and Bruce Edmonds. A conversational intelligent tutoring system to automatically predict learning styles. *Computers & Education*, 59(1):95–109, 2012.

[130] Joey J Lee, Pinar Ceyhan, William Jordan-Cooley, and Woonhee Sung. GREENIFY: A real-world action game for climate change education. *Simulation & Gaming*, 44(2-3):349–365, 2013.

[131] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710, 1966.

[132] Rich Ling and Chih-Hui Lai. Microcoordination 2.0: Social coordination in the age of smartphones and messaging apps. *Journal of Communication*, 66(5):834–856, 2016.

[133] Ian J Livingston, Carl Gutwin, Regan L Mandryk, and Max Birk. How players value their characters in world of warcraft. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1333–1343, 2014.

[134] Richard J Loewenstein. An office mental status examination for complex chronic dissociative symptoms and multiple personality disorder. *Psychiatric Clinics*, 14(3):567–604, 1991.

[135] Gale M Lucas, Albert Rizzo, Jonathan Gratch, Stefan Scherer, Giota Stratou, Jill Boberg, and Louis-Philippe Morency. Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI*, 4:51, 2017.

[136] Ewa Luger and Abigail Sellen. 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*, pages 5286–5297. ACM, 2016.

[137] Alwin Luttmann, Klaus-Helmut Schmidt, and Matthias Jäger. Working conditions, muscular activity and complaints of office workers. *International Journal of Industrial Ergonomics*, 40(5):549–559, 2010.

[138] V K Manam and Alexander Quinn. Wingit: Efficient refinement of unclear task instructions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 6, 2018.

[139] Andrew Mao, Ece Kamar, and Eric Horvitz. Why stop now? predicting worker engagement in online crowdsourcing. In *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing*, pages 103–111. AAAI, 2013.

[140] Catherine C Marshall and Frank M Shipman. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 234–243. ACM, 2013.

[141] David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 224–235, 2014.

[142] Elaine Massung, David Coyle, Kirsten F Cater, Marc Jay, and Chris Preist. Using crowdsourcing to support pro-environmental community activism. In *Proceedings of the SIGCHI Conference on human factors in Computing systems*, pages 371–380, 2013.

[143] Panagiotis Mavridis, David Gross-Amblard, and Zoltán Miklós. Using Hierarchical Skills for Optimized Task Assignment in Knowledge-Intensive Crowdsourcing. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 843–853, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

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

[145] Edward McAuley, Terry Duncan, and Vance V Tammen. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research quarterly for exercise and sport*, 60(1):48–58, 1989.

[146] Michael P McCreery, S Kathleen Krach, Peter G Schrader, and Randy Boone. Defining the virtual self: Personality, behavior, and the psychology of embodiment. *Computers in Human Behavior*, 28(3):976–983, 2012.

[147] James L McGaugh. Memory–a century of consolidation. *Science*, 287(5451):248–251, 2000.

[148] Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. Taking a HIT: Designing around rejection, mistrust, risk, and workers' experiences in Amazon Mechanical Turk. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2271–2282, 2016.

[149] Catherine Houdek Middendorf and Therese Hoff Macan. Note-taking in the employment interview: Effects on recall and judgments. *Journal of Applied Psychology*, 87(2):293, 2002.

[150] Adam S Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine*, 176(5):619–625, 2016.

[151] Andrew G Miner and Theresa M Glomb. State mood, task performance, and behavior at work: A within-persons approach. *Organizational Behavior and Human Decision Processes*, 112(1):43–57, 2010.

[152] Jesus Montero-Marin, Javier Garcia-Campayo, Mari Cruz Pérez-Yus, Edurne Zabaleta-del Olmo, and Pim Cuijpers. Meditation techniques v. relaxation therapies when treating anxiety: A meta-analytic review. *Psychological medicine*, 49(13):2118–2133, 2019.

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

[154] Benedikt Morschheuser, Juho Hamari, Jonna Koivisto, and Alexander Maedche. Gamified crowdsourcing: Conceptualization, literature review, and future agenda. *International Journal of Human-Computer Studies*, 106:26–43, 2017.

[155] Georg Elias Müller and Alfons Pilzecker. *Experimentelle beiträge zur lehre vom gedächtniss*, volume 1. JA Barth, 1900.

[156] Sean A Munson, Karina Kervin, and Lionel P Robert Jr. Monitoring email to indicate project team performance and mutual attraction. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 542–549, 2014.

[157] Jaap M J Murre and Joeri Dros. Replication and analysis of Ebbinghaus' forgetting curve. *PloS one*, 10(7), 2015.

[158] K Murrell. *Ergonomics: Man in his working environment.* Springer Science & Business Media, 2012.

[159] Prayag Narula, Philipp Gutheim, David Rolnitzky, Anand Kulkarni, and Bjoern Hartmann. MobileWorks: A Mobile Crowdsourcing Platform for Workers at the Bottom of the Pyramid. *Human Computation*, 11(11):45, 2011.

[160] Carman Neustaedter and Elena A Fedorovskaya. Presenting identity in a virtual world through avatar appearances. In *Graphics Interface*, pages 183–190, 2009.

[161] Dong Nguyen, Noah A Smith, and Carolyn P Rosé. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, pages 115–123. Association for Computational Linguistics, 2011.

[162] Catarina Nordander, Gert-Åke Hansson, Lars Rylander, Paul Asterland, Jeannette Unge BystrÖm, Kerstina Ohlsson, Istvan Balogh, and Staffan Skerfving. Muscular rest and gap frequency as EMG measures of physical exposure: the impact of work tasks and individual related. *Ergonomics*, 43(11):1904–1919, 2000.

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

[164] Heather L O'Brien, Paul Cairns, and Mark Hall. 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*, 112:28–39, 2018.

[165] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, 2017.

[166] Reinhard Pekrun, Thomas Goetz, Anne C Frenzel, Petra Barchfeld, and Raymond P Perry. Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary educational psychology*, 36(1):36–48, 2011.

[167] Lisa M Perez, Jeremy Jones, David R Englert, and Daniel Sachau. Secondary traumatic stress and burnout among law enforcement investigators exposed to disturbing media images. *Journal of Police and Criminal Psychology*, 25(2):113–124, 2010.

[168] Pierre Philip, Jean-Arthur Micoulaud-Franchi, Patricia Sagaspe, Etienne De Sevin, Jérôme Olive, Stéphanie Bioulac, and Alain Sauteraud. Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. *Scientific reports*, 7(1):1–7, 2017.

[169] Sihang Qiu, Alessandro Bozzon, Max V Birk, and Ujwal Gadiraju. Using Worker Avatars to Improve Microtask Crowdsourcing. In *Proceedings of the ACM on Human-Computer Interaction*, pages 1–28. ACM New York, NY, USA, 2021.

[170] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Estimating Conversational Styles in Conversational Microtask Crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–23, 2020.

[171] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Improving worker engagement through conversational microtask crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[172] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Just the Right Mood for HIT! In *International Conference on Web Engineering*, pages 381–396. Springer, 2020.

[173] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. TickTalkTurk: Conversational Crowdsourcing Made Easy. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, pages 1–5, 2020.

[174] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Towards memorable information retrieval. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pages 69–76, 2020.

[175] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Understanding Worker Health on Crowdsourcing Platforms. In submission, 2021.

[176] Sihang Qiu, Achilleas Psyllidis, Alessandro Bozzon, and Geert-Jan Houben. Crowd-Mapping Urban Objects from Street-Level Imagery. In *The World Wide Web Conference*, pages 1521–1531. ACM, 2019.

[177] Priyanga Ranasinghe, Yashasvi S Perera, Dilusha A Lamabadusuriya, Supun Kulatunga, Naveen Jayawardana, Senaka Rajapakse, and Prasad Katulanda. Work related complaints of neck, shoulder and arm among computer office workers: a cross-sectional evaluation of prevalence and risk factors in a developing country. *Environmental Health*, 10(1):1–9, 2011.

[178] Hyekyun Rhee, James Allen, Jennifer Mammen, and Mary Swift. Mobile phone-based asthma self-management aid for adolescents (mASMAA): a feasibility study. *Patient preference and adherence*, 8:63, 2014.

[179] Bradley Rhodes and Thad Starner. Remembrance Agent: A continuously running automated information retrieval system. In *The Proceedings of The First International Conference on The Practical Application Of Intelligent Agents and Multi Agent Technology*, pages 487–495, 1996.

[180] Christopher Riederer, Jake M Hofman, and Daniel G Goldstein. To put that in perspective: Generating analogies that make numbers easier to understand. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 548. ACM, 2018.

[181] Martin J Riedl, Gina M Masullo, and Kelsey N Whipple. The downsides of digital labor: Exploring the toll incivility takes on online comment moderators. *Computers in Human Behavior*, 107:106262, 2020.

[182] Markus Rokicki, Sergiu Chelaru, Sergej Zerr, and Stefan Siersdorfer. Competitive game designs for improving the cost effectiveness of crowdsourcing. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1469–1478. ACM, 2014.

[183] Markus Rokicki, Sergej Zerr, and Stefan Siersdorfer. Groupsourcing: Team competition designs for crowdsourcing. In *Proceedings of the 24th international conference on world wide web*, pages 906–915, 2015.

[184] Richard M Ryan, C Scott Rigby, and Andrew Przybylski. The motivational pull of video games: A self-determination theory approach. *Motivation and emotion*, 30(4):344–360, 2006.

[185] Rufat Rzayev, Tilman Dingler, and Niels Henze. ReflectiveDiary: Fostering Human Memory Through Activity Summaries Created from Implicit Data Collection. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*, MUM 2018, pages 285–291, New York, NY, USA, 2018. ACM.

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

[187] Susumu Saito, Chun-Wei Chiang, Saiph Savage, Teppei Nakano, Tetsunori Kobayashi, and Jeffrey P Bigham. TurkScanner: Predicting the hourly wage of microtasks. In *The World Wide Web Conference*, pages 3187–3193, 2019.

[188] M Rafael Salaberry. L2 morphosyntactic development in text-based computer-mediated communication. *Computer Assisted Language Learning*, 13(1):5–27, 2000.

[189] Niloufar Salehi, Lilly C Irani, Michael S Bernstein, Ali Alkhatib, Eva Ogbe, and Kristy Milland. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 1621–1630, 2015.

[190] Shruti Sannon and Dan Cosley. Privacy, Power, and Invisible Labor on Amazon Mechanical Turk. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[191] John W Satzinger and Lorne Olfman. User interface consistency across end-user applications: the effects on mental models. *Journal of Management Information Systems*, 14(4):167–193, 1998.

[192] Saiph Savage, Chun Wei Chiang, Susumu Saito, Carlos Toxtli, and Jeffrey Bigham. Becoming the Super Turker: Increasing Wages via a Strategy from High Earning Workers. In *Proceedings of The Web Conference 2020*, pages 1241–1252, 2020.

[193] Saiph Savage, Andres Monroy-Hernandez, and Tobias Höllerer. Botivist: Calling volunteers to action using online bots. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 813–822. ACM, 2016.

[194] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs.*, pages 199–205, 2006.

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

[196] Jack Sidnell and Tanya Stivers. *The handbook of conversation analysis*, volume 121. John Wiley & Sons, 2012.

[197] M Silberman and Lilly Irani. Operating an employer reputation system: lessons from Turkopticon, 2008-2015. *Comparative Labor Law & Policy Journal, Forthcoming*, 2016.

[198] Rachel N Simons, Danna Gurari, and Kenneth R Fleischmann. "I Hope This Is Helpful" Understanding Crowdworkers' Challenges and Motivations for an Image Description Task. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26, 2020.

[199] Norman J Slamecka and Peter Graf. The generation effect: Delineation of a phenomenon. *Journal of experimental Psychology: Human learning and Memory*, 4(6):592, 1978.

[200] Virpi Slotte and Kirsti Lonka. Review and process effects of spontaneous note-taking on text comprehension. *Contemporary Educational Psychology*, 24(1):1–20, 1999.

[201] Jessica Smith. The Messaging Apps Report: How brands, businesses, and publishers can capitalize on the rising tide of messaging platforms. `https://www.businessinsider.com/messaging-apps-report-2018-4`, 2018.

[202] Stacy L Smith and Edward Donnerstein. Harmful effects of exposure to media violence: Learning of aggression, emotional desensitization, and fear. In *Human aggression*, pages 167–202. Elsevier, 1998.

[203] Donggil Song, Eun Young Oh, and Marilyn Rice. Interacting with a conversational agent system for educational purposes in online courses. In *2017 10th international conference on human system interactions (HSI)*, pages 78–82. IEEE, 2017.

[204] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. The Psychological Well-Being of Content Moderators. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14, 2021.

[205] Franklin Stein. Occupational stress, relaxation therapies, exercise and biofeedback. *Work*, 17(3):235–245, 2001.

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

[207] Anselm L Strauss. *Qualitative analysis for social scientists.* Cambridge university press, 1987.

[208] Ryo Suzuki, Niloufar Salehi, Michelle S Lam, Juan C Marroquin, and Michael S Bernstein. Atelier: Repurposing expert crowdsourcing tasks as micro-internships. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2645–2656, 2016.

[209] Hiroki Tanaka, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PloS one*, 12(8):e0182151, 2017.

[210] Deborah Tannen. Conversational style. *Psycholinguistic models of production*, pages 251–267, 1987.

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

[212] Syed H Tariq, Nina Tumosa, John T Chibnall, Mitchell H Perry III, and John E Morley. Comparison of the Saint Louis University mental status examination and the mini-mental state examination for detecting dementia and mild neurocognitive disorder - a pilot study. *The American journal of geriatric psychiatry*, 14(11):900–910, 2006.

[213] Yla R Tausczik and James W Pennebaker. Improving teamwork using real-time language feedback. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 459–468, 2013.

[214] Jaime Teevan, Brent Hecht, Sonia Jaffe, and Eds. The New Future of Work: Research from Microsoft on the Impact of the Pandemic on Work Practices. Technical report, Microsoft, 2021.

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

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

[217] P A Totterdell and Karen Niven. *Workplace moods and emotions: A review of research.* Createspace Independent Publishing, 2014.

[218] Carlos Toxtli, Joel Chan, Walter S Lasecki, and Saiph Savage. Enabling Expert Critique with Chatbots and Micro Guidance. In *Collective Intelligence 2018*, page 4. ACM, 2018.

[219] Nam Khanh Tran, Andrea Ceroni, Nattiya Kanhabua, and Claudia Niederée. Back to the past: Supporting interpretations of forgotten stories by time-aware re-contextualization. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 339–348. ACM, 2015.

[220] Sabine Trepte and Leonard Reinecke. Avatar creation and video game enjoyment: Effects of life-satisfaction, game competitiveness, and identification with the avatar. *Journal of Media Psychology: Theories, Methods, and Applications*, 22(4), 2010.

[221] Kristen Vaccaro, Tanvi Agarwalla, Sunaya Shivakumar, and Ranjitha Kumar. Designing the Future of Personal Fashion. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 627:1—-627:11, New York, NY, USA, 2018. ACM.

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

[223] Philippe Verduyn, Iven Van Mechelen, and Francis Tuerlinckx. The relation between event processing and the duration of emotional experience. *Emotion*, 11(1):20, 2011.

[224] Andrew Vickers and Catherine Zollman. Hypnosis and relaxation therapies. *Bmj*, 319(7221):1346–1349, 1999.

[225] Peter Vorderer, Christoph Klimmt, and Ute Ritterfeld. Enjoyment: At the heart of media entertainment. *Communication theory*, 14(4):388–408, 2004.

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

[227] J E Ware, Kristin K Snow, Mark Kosinski, and Barbara Gandek. SF-36 health survey. *Manual and interpretation guide. Boston: The Health Institute, New England Medical Center*, pages 10–16, 1993.

[228] John E Ware Jr. SF-36 health survey update. *Spine*, 25(24):3130–3139, 2000.

[229] Mark E Whiting, Grant Hugh, and Michael S Bernstein. Fair work: Crowd work minimum wage with one line of code. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 197–206, 2019.

[230] Alex C Williams, Gloria Mark, Kristy Milland, Edward Lank, and Edith Law. The Perpetual Work Life of Crowdworkers: How Tooling Practices Increase Fragmentation in Crowdwork. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–28, 2019.

[231] E H C Woo, Peter White, and C W K Lai. Ergonomics standards and guidelines for computer workstation design and the impact on users' health–a review. *Ergonomics*, 59(3):464–475, 2016.

[232] World Health Organization. The World Health Report 2001: Mental health: new understanding, new hope. Technical report, World Health Organization, 2001.

[233] Thomas A Wright and Russell Cropanzano. The happy/productive worker thesis revisited. In *Research in personnel and human resources management*, pages 269–307. Emerald Group Publishing Limited, 2007.

[234] Meng-Han Wu and Alexander Quinn. Confusing the crowd: Task instruction quality on amazon mechanical turk. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 5, 2017.

[235] Luyan Xu, Xuan Zhou, and Ujwal Gadiraju. Revealing the Role of User Moods in Struggling Search Tasks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1249–1252. ACM, 2019.

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

[237] Jie Yang, Ke Tao, Alessandro Bozzon, and Geert-Jan Houben. Sparrows and Owls: Characterisation of Expert Behaviour in StackOverflow. In Vania Dimitrova, Tsvi Kuflik, David Chin, Francesco Ricci, Peter Dolog, and Geert-Jan Houben, editors, *User Modeling, Adaptation, and Personalization: 22nd International Conference, UMAP 2014, Aalborg, Denmark, July 7-11, 2014. Proceedings*, pages 266–277. Springer International Publishing, Cham, 2014.

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

[239] Ji Won You and Myunghee Kang. The role of academic emotions in the relationship between perceived academic control and self-regulated learning in online learning. *Computers & Education*, 77:125–133, 2014.

[240] Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. Predicting user knowledge gain in informational search sessions. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 75–84, 2018.

[241] John M Zelenski, Steven A Murphy, and David A Jenkins. The happy-productive worker thesis revisited. *Journal of Happiness Studies*, 9(4):521–537, 2008.

[242] Ying Zhang, Xianghua Ding, and Ning Gu. Understanding fatigue and its impact in crowdsourcing. In *2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD))*, pages 57–62. IEEE, 2018.

[243] Yudian Zheng, Jiannan Wang, Guoliang Li, Reynold Cheng, and Jianhua Feng. QASCA: A quality-aware task assignment system for crowdsourcing applications. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, pages 1031–1046, 2015.

[244] Mengdie Zhuang and Ujwal Gadiraju. In What Mood Are You Today? An Analysis of Crowd Workers' Mood, Performance and Engagement. In *Proceedings of the 10th ACM Conference on Web Science*, pages 373–382, 2019.

[245] Victor W Zue and James R Glass. Conversational interfaces: Advances and challenges. *Proceedings of the IEEE*, 88(8):1166–1180, 2000.

[246] Darius Zumstein and Sophie Hundertmark. Chatbots–An Interactive Technology for personalized communication, transactions and services. *IADIS International Journal on WWW/Internet*, 15(1), 2017.

# List of Figures

# List of Tables

# Curriculum Vitæ

Sihang Qiu is a PhD student at the Web Information Systems group of the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, supervised by Geert-Jan Houben, Alessandro Bozzon, and Ujwal Gadiraju. His PhD work focuses on the intersection of conversational systems and crowd computing.

Sihang Qiu was born in Zhejiang, China on May 27, 1993. He obtained his Bachelor degree and Master degree from National University of Defense Technology in 2015 and 2017 respectively.

## Publications

**Conference/Journal Publications:**

Sihang Qiu, Alessandro Bozzon, Max V. Birk, Ujwal Gadiraju. Using Worker Avatars to Improve Microtask Crowdsourcing. Proceedings of the ACM on Human-Computer Interaction (CSCW), pp. 1-28, 2021.

Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon. Estimating Conversational Styles in Conversational Microtask Crowdsourcing. Proceedings of the ACM on Human-Computer Interaction (CSCW), vol. 4, pp. 1-23, 2020.

Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon. Improving Worker Engagement Through Conversational Microtask Crowdsourcing. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1-12, 2020.

Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon. Just the Right Mood for HIT! Analyzing the Role of Worker Moods in Conversational Microtask Crowdsourcing. International Conference on Web Engineering, pp. 381-396, 2020.

Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon. Towards Memorable Information Retrieval. Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval, pp. 69-76, 2020.

Sihang Qiu, Achilleas Psyllidis, Alessandro Bozzon, Geert-Jan Houben. Crowd-Mapping Urban Objects from Street-Level Imagery. The World Wide Web Conference, pp. 1521-1531, 2019.

Ioannis Petros Samiotis, Sihang Qiu, Christoph Lofi, Jie Yang, Ujwal Gadiraju and Alessandro Bozzon. Exploring the Music Perception Skills of Crowd Workers. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 2021.

Tom Edixhoven, Sihang Qiu, Lucie Kuiper, Olivier Dikken, Gwennan Smitskamp, Ujwal Gadiraju. Improving Reactions to Rejection in Crowdsourcing Through Self-Reflection. 13th ACM Web Science Conference, 2021.

Gerard van Alphen, Sihang Qiu, Alessandro Bozzon, Geert-Jan Houben. Analyzing Workers Performance in Online Mapping Tasks Across Web, Mobile, and Virtual Reality Platforms. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 8, pp. 141-149, 2020.

Shahin Sharifi Noorian, Sihang Qiu, Achilleas Psyllidis, Alessandro Bozzon, Geert-Jan Houben. Detecting, Classifying, and Mapping Retail Storefronts Using Street-level Imagery. Proceedings of the 2020 International Conference on Multimedia Retrieval, pp. 495-501, 2020.

Ioannis Petros Samiotis, Sihang Qiu, Andrea Mauri, Cynthia Liem, Christoph Lofi, Alessandro Bozzon. Microtask Crowdsourcing for Music Score Transcriptions: An Experiment with Error Detection. Proceedings of the 21st Conference of the International Society for Music Information Retrieval, pp. 1-7, 2020.

Shabnam Najafian, Daniel Herzog, Sihang Qiu, Oana Inel, Nava Tintarev. You do not decide for me! Evaluating Explainable Group Aggregation Strategies for Tourism. Proceedings of the 31st ACM Conference on Hypertext and Social Media, pp. 187-196, 2020.

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

**Demo/Workshop/Work-in-Progress Publications:**

Willem van der Maden, Sihang Qiu, James Lomas, Ujwal Gadiraju. Context-Sensitive Assessments of Human Wellbeing. CHI 2021 Workshop: Designing for New Forms of Vulnerability, 2021.

Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon. TickTalkTurk: Conversational Crowdsourcing Made Easy. Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing, pp.53-57, 2020.

Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon, Geert-Jan Houben. Conversational Crowdsourcing. Proceedings of the Crowd Science Workshop: Remoteness, Fairness, and Mechanisms as Challenges of Data Supply by Humans for Automation co-located with 34th Conference on Neural Information Processing Systems, vol. 2736, pp. 1-6, 2020.

Sihang Qiu, Alessandro Bozzon, Ujwal Gadiraju. Conversational Interfaces for Search As Learning. CIKMW2020: Proceeding of the CIKM 2020 Workshops, vol. 2699, pp. 1-4, 2020.

Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon. Remote Work Aided by Conversational Agents. New Future of Work, pp. 1-5, 2020.

Sihang Qiu, Alessandro Bozzon, Geert-Jan Houben. VirtualCrowd: A Simulation Platform for Microtask Crowdsourcing Campaigns. Companion Proceedings of the Web Conference 2020, pp. 222-225, 2020.

Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon. Understanding Conversational Style in Conversational Microtask Crowdsourcing. Work in Progress and Demo track of HCOMP 2019, pp. 1-3, 2019.

Nava Tintarev, Emily Sullivan, Dror Guldin, Sihang Qiu, Daan Odjik. Same, Same, but Different: Algorithmic Diversification of Viewpoints in News. Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, pp. 7-13, 2018.

# SIKS Dissertation Series

Since 1998, all dissertations written by Ph.D. students who have conducted their research under auspices of a senior research fellow of the SIKS research school are published in the SIKS Dissertation Series.

2011 01 Botond Cseke (RUN), Variational Algorithms for Bayesian Inference in Latent Gaussian Models

02 Nick Tinnemeier (UU), Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language

03 Jan Martijn van der Werf (TUE), Compositional Design and Verification of Component-Based Information Systems

04 Hado van Hasselt (UU), Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference

05 Bas van der Raadt (VU), Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.

06 Yiwen Wang (TUE), Semantically-Enhanced Recommendations in Cultural Heritage

07 Yujia Cao (UT), Multimodal Information Presentation for High Load Human Computer Interaction

08 Nieske Vergunst (UU), BDI-based Generation of Robust Task-Oriented Dialogues

09 Tim de Jong (OU), Contextualised Mobile Media for Learning

10 Bart Bogaert (UvT), Cloud Content Contention

11 Dhaval Vyas (UT), Designing for Awareness: An Experience-focused HCI Perspective

12 Carmen Bratosin (TUE), Grid Architecture for Distributed Process Mining

13 Xiaoyu Mao (UvT), Airport under Control. Multiagent Scheduling for Airport Ground Handling

14 Milan Lovric (EUR), Behavioral Finance and Agent-Based Artificial Markets

15 Marijn Koolen (UvA), The Meaning of Structure: the Value of Link Evidence for Information Retrieval

16 Maarten Schadd (UM), Selective Search in Games of Different Complexity

17 Jiyin He (UVA), Exploring Topic Structure: Coherence, Diversity and Relatedness

18 Mark Ponsen (UM), Strategic Decision-Making in complex games

19 Ellen Rusman (OU), The Mind's Eye on Personal Profiles

20 Qing Gu (VU), Guiding service-oriented software engineering - A view-based approach

21 Linda Terlouw (TUD), Modularization and Specification of Service-Oriented Systems

22 Junte Zhang (UVA), System Evaluation of Archival Description and Access

23 Wouter Weerkamp (UVA), Finding People and their Utterances in Social Media

24 Herwin van Welbergen (UT), Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior

25 Syed Waqar ul Qounain Jaffry (VU), Analysis and Validation of Models for Trust Dynamics

26 Matthijs Aart Pontier (VU), Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots

27 Aniel Bhulai (VU), Dynamic website optimization through autonomous management of design patterns

28 Rianne Kaptein (UVA), Effective Focused Retrieval by Exploiting Query Context and Document Structure

29 Faisal Kamiran (TUE), Discrimination-aware Classification

30 Egon van den Broek (UT), Affective Signal Processing (ASP): Unraveling the mystery of emotions

31 Ludo Waltman (EUR), Computational and Game-Theoretic Approaches for Modeling Bounded Rationality

32 Nees-Jan van Eck (EUR), Methodological Advances in Bibliometric Mapping of Science

33 Tom van der Weide (UU), Arguing to Motivate Decisions

34 Paolo Turrini (UU), Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations

35 Maaike Harbers (UU), Explaining Agent Behavior in Virtual Training

36 Erik van der Spek (UU), Experiments in serious game design: a cognitive approach

37 Adriana Burlutiu (RUN), Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference

38 Nyree Lemmens (UM), Bee-inspired Distributed Optimization

39 Joost Westra (UU), Organizing Adaptation using Agents in Serious Games

40 Viktor Clerc (VU), Architectural Knowledge Management in Global Software Development

41 Luan Ibraimi (UT), Cryptographically Enforced Distributed Data Access Control

42 Michal Sindlar (UU), Explaining Behavior through Mental State Attribution

43 Henk van der Schuur (UU), Process Improvement through Software Operation Knowledge

44 Boris Reuderink (UT), Robust Brain-Computer Interfaces

45 Herman Stehouwer (UvT), Statistical Language Models for Alternative Sequence Selection

46 Beibei Hu (TUD), Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work

47 Azizi Bin Ab Aziz (VU), Exploring Computational Models for Intelligent Support of Persons with Depression

48 Mark Ter Maat (UT), Response Selection and Turn-taking for a Sensitive Artificial Listening Agent

49 Andreea Niculescu (UT), Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

---

2012 01 Terry Kakeeto (UvT), Relationship Marketing for SMEs in Uganda

02 Muhammad Umair (VU), Adaptivity, emotion, and Rationality in Human and Ambient Agent Models

03 Adam Vanya (VU), Supporting Architecture Evolution by Mining Software Repositories

04 Jurriaan Souer (UU), Development of Content Management System-based Web Applications

05 Marijn Plomp (UU), Maturing Interorganisational Information Systems

06 Wolfgang Reinhardt (OU), Awareness Support for Knowledge Workers in Research Networks

07 Rianne van Lambalgen (VU), When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions

08 Gerben de Vries (UVA), Kernel Methods for Vessel Trajectories

09 Ricardo Neisse (UT), Trust and Privacy Management Support for Context-Aware Service Platforms

10 David Smits (TUE), Towards a Generic Distributed Adaptive Hypermedia Environment

11 J.C.B. Rantham Prabhakara (TUE), Process Mining in the Large: Preprocessing, Discovery, and Diagnostics

12 Kees van der Sluijs (TUE), Model Driven Design and Data Integration in Semantic Web Information Systems

13 Suleman Shahid (UvT), Fun and Face: Exploring non-verbal expressions of emotion during playful interactions

14 Evgeny Knutov (TUE), Generic Adaptation Framework for Unifying Adaptive Web-based Systems

15 Natalie van der Wal (VU), Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.

16 Fiemke Both (VU), Helping people by understanding them - Ambient Agents supporting task execution and depression treatment

17 Amal Elgammal (UvT), Towards a Comprehensive Framework for Business Process Compliance

18 Eltjo Poort (VU), Improving Solution Architecting Practices

19 Helen Schonenberg (TUE), What's Next? Operational Support for Business Process Execution

20 Ali Bahramisharif (RUN), Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing

21 Roberto Cornacchia (TUD), Querying Sparse Matrices for Information Retrieval

22 Thijs Vis (UvT), Intelligence, politie en veiligheidsdienst: verenigbare grootheden?

23 Christian Muehl (UT), Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction

24 Laurens van der Werff (UT), Evaluation of Noisy Transcripts for Spoken Document Retrieval

25 Silja Eckartz (UT), Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application

26 Emile de Maat (UVA), Making Sense of Legal Text

27 Hayrettin Gurkok (UT), Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games

28 Nancy Pascall (UvT), Engendering Technology Empowering Women

29 Almer Tigelaar (UT), Peer-to-Peer Information Retrieval

30 Alina Pommeranz (TUD), Designing Human-Centered Systems for Reflective Decision Making

31 Emily Bagarukayo (RUN), A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure

32 Wietske Visser (TUD), Qualitative multi-criteria preference representation and reasoning
33 Rory Sie (OUN), Coalitions in Cooperation Networks (COCOON)
34 Pavol Jancura (RUN), Evolutionary analysis in PPI networks and applications
35 Evert Haasdijk (VU), Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics
36 Denis Ssebugwawo (RUN), Analysis and Evaluation of Collaborative Modeling Processes
37 Agnes Nakakawa (RUN), A Collaboration Process for Enterprise Architecture Creation
38 Selmar Smit (VU), Parameter Tuning and Scientific Testing in Evolutionary Algorithms
39 Hassan Fatemi (UT), Risk-aware design of value and coordination networks
40 Agus Gunawan (UvT), Information Access for SMEs in Indonesia
41 Sebastian Kelle (OU), Game Design Patterns for Learning
42 Dominique Verpoorten (OU), Reflection Amplifiers in self-regulated Learning
43 Withdrawn
44 Anna Tordai (VU), On Combining Alignment Techniques
45 Benedikt Kratz (UvT), A Model and Language for Business-aware Transactions
46 Simon Carter (UVA), Exploration and Exploitation of Multilingual Data for Statistical Machine Translation
47 Manos Tsagkias (UVA), Mining Social Media: Tracking Content and Predicting Behavior
48 Jorn Bakker (TUE), Handling Abrupt Changes in Evolving Time-series Data
49 Michael Kaisers (UM), Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions
50 Steven van Kervel (TUD), Ontologogy driven Enterprise Information Systems Engineering
51 Jeroen de Jong (TUD), Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

---

2013 01 Viorel Milea (EUR), News Analytics for Financial Decision Support
02 Erietta Liarou (CWI), MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing
03 Szymon Klarman (VU), Reasoning with Contexts in Description Logics
04 Chetan Yadati (TUD), Coordinating autonomous planning and scheduling
05 Dulce Pumareja (UT), Groupware Requirements Evolutions Patterns
06 Romulo Goncalves (CWI), The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience
07 Giel van Lankveld (UvT), Quantifying Individual Player Differences
08 Robbert-Jan Merk (VU), Making enemies: cognitive modeling for opponent agents in fighter pilot simulators
09 Fabio Gori (RUN), Metagenomic Data Analysis: Computational Methods and Applications
10 Jeewanie Jayasinghe Arachchige (UvT), A Unified Modeling Framework for Service Design.
11 Evangelos Pournaras (TUD), Multi-level Reconfigurable Self-organization in Overlay Services
12 Marian Razavian (VU), Knowledge-driven Migration to Services
13 Mohammad Safiri (UT), Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly
14 Jafar Tanha (UVA), Ensemble Approaches to Semi-Supervised Learning Learning
15 Daniel Hennes (UM), Multiagent Learning - Dynamic Games and Applications
16 Eric Kok (UU), Exploring the practical benefits of argumentation in multi-agent deliberation
17 Koen Kok (VU), The PowerMatcher: Smart Coordination for the Smart Electricity Grid
18 Jeroen Janssens (UvT), Outlier Selection and One-Class Classification
19 Renze Steenhuizen (TUD), Coordinated Multi-Agent Planning and Scheduling
20 Katja Hofmann (UvA), Fast and Reliable Online Learning to Rank for Information Retrieval
21 Sander Wubben (UvT), Text-to-text generation by monolingual machine translation
22 Tom Claassen (RUN), Causal Discovery and Logic
23 Patricio de Alencar Silva (UvT), Value Activity Monitoring
24 Haitham Bou Ammar (UM), Automated Transfer in Reinforcement Learning
25 Agnieszka Anna Latoszek-Berendsen (UM), Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System
26 Alireza Zarghami (UT), Architectural Support for Dynamic Homecare Service Provisioning
27 Mohammad Huq (UT), Inference-based Framework Managing Data Provenance
28 Frans van der Sluis (UT), When Complexity becomes Interesting: An Inquiry into the Information eXperience
29 Iwan de Kok (UT), Listening Heads
30 Joyce Nakatumba (TUE), Resource-Aware Business Process Management: Analysis and Support

31 Dinh Khoa Nguyen (UvT), Blueprint Model and Language for Engineering Cloud Applications

32 Kamakshi Rajagopal (OUN), Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development

33 Qi Gao (TUD), User Modeling and Personalization in the Microblogging Sphere

34 Kien Tjin-Kam-Jet (UT), Distributed Deep Web Search

35 Abdallah El Ali (UvA), Minimal Mobile Human Computer Interaction

36 Than Lam Hoang (TUe), Pattern Mining in Data Streams

37 Dirk Börner (OUN), Ambient Learning Displays

38 Eelco den Heijer (VU), Autonomous Evolutionary Art

39 Joop de Jong (TUD), A Method for Enterprise Ontology based Design of Enterprise Information Systems

40 Pim Nijssen (UM), Monte-Carlo Tree Search for Multi-Player Games

41 Jochem Liem (UVA), Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning

42 Léon Planken (TUD), Algorithms for Simple Temporal Reasoning

43 Marc Bron (UVA), Exploration and Contextualization through Interaction and Concepts

2014 01 Nicola Barile (UU), Studies in Learning Monotone Models from Data

02 Fiona Tuliyano (RUN), Combining System Dynamics with a Domain Modeling Method

03 Sergio Raul Duarte Torres (UT), Information Retrieval for Children: Search Behavior and Solutions

04 Hanna Jochmann-Mannak (UT), Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation

05 Jurriaan van Reijsen (UU), Knowledge Perspectives on Advancing Dynamic Capability

06 Damian Tamburri (VU), Supporting Networked Software Development

07 Arya Adriansyah (TUE), Aligning Observed and Modeled Behavior

08 Samur Araujo (TUD), Data Integration over Distributed and Heterogeneous Data Endpoints

09 Philip Jackson (UvT), Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language

10 Ivan Salvador Razo Zapata (VU), Service Value Networks

11 Janneke van der Zwaan (TUD), An Empathic Virtual Buddy for Social Support

12 Willem van Willigen (VU), Look Ma, No Hands: Aspects of Autonomous Vehicle Control

13 Arlette van Wissen (VU), Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains

14 Yangyang Shi (TUD), Language Models With Meta-information

15 Natalya Mogles (VU), Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare

16 Krystyna Milian (VU), Supporting trial recruitment and design by automatically interpreting eligibility criteria

17 Kathrin Dentler (VU), Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability

18 Mattijs Ghijsen (UVA), Methods and Models for the Design and Study of Dynamic Agent Organizations

19 Vinicius Ramos (TUE), Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support

20 Mena Habib (UT), Named Entity Extraction and Disambiguation for Informal Text: The Missing Link

21 Kassidy Clark (TUD), Negotiation and Monitoring in Open Environments

22 Marieke Peeters (UU), Personalized Educational Games - Developing agent-supported scenario-based training

23 Eleftherios Sidirourgos (UvA/CWI), Space Efficient Indexes for the Big Data Era

24 Davide Ceolin (VU), Trusting Semi-structured Web Data

25 Martijn Lappenschaar (RUN), New network models for the analysis of disease interaction

26 Tim Baarslag (TUD), What to Bid and When to Stop

27 Rui Jorge Almeida (EUR), Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty

28 Anna Chmielowiec (VU), Decentralized k-Clique Matching

29 Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software

30 Peter de Cock (UvT), Anticipating Criminal Behaviour

31 Leo van Moergestel (UU), Agent Technology in Agile Multiparallel Manufacturing and Product Support

32 Naser Ayat (UvA), On Entity Resolution in Probabilistic Data

33 Tesfa Tegegne (RUN), Service Discovery in eHealth

34 Christina Manteli (VU), The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems.

35 Joost van Ooijen (UU), Cognitive Agents in Virtual Worlds: A Middleware Design Approach

36 Joos Buijs (TUE), Flexible Evolutionary Algorithms for Mining Structured Process Models
37 Maral Dadvar (UT), Experts and Machines United Against Cyberbullying
38 Danny Plass-Oude Bos (UT), Making brain-computer interfaces better: improving usability through post-processing.
39 Jasmina Maric (UvT), Web Communities, Immigration, and Social Capital
40 Walter Omona (RUN), A Framework for Knowledge Management Using ICT in Higher Education
41 Frederic Hogenboom (EUR), Automated Detection of Financial Events in News Text
42 Carsten Eijckhof (CWI/TUD), Contextual Multidimensional Relevance Models
43 Kevin Vlaanderen (UU), Supporting Process Improvement using Method Increments
44 Paulien Meesters (UvT), Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.
45 Birgit Schmitz (OUN), Mobile Games for Learning: A Pattern-Based Approach
46 Ke Tao (TUD), Social Web Data Analytics: Relevance, Redundancy, Diversity
47 Shangsong Liang (UVA), Fusion and Diversification in Information Retrieval

---

2015 01 Niels Netten (UvA), Machine Learning for Relevance of Information in Crisis Response
02 Faiza Bukhsh (UvT), Smart auditing: Innovative Compliance Checking in Customs Controls
03 Twan van Laarhoven (RUN), Machine learning for network data
04 Howard Spoelstra (OUN), Collaborations in Open Learning Environments
05 Christoph Bösch (UT), Cryptographically Enforced Search Pattern Hiding
06 Farideh Heidari (TUD), Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes
07 Maria-Hendrike Peetz (UvA), Time-Aware Online Reputation Analysis
08 Jie Jiang (TUD), Organizational Compliance: An agent-based model for designing and evaluating organizational interactions
09 Randy Klaassen (UT), HCI Perspectives on Behavior Change Support Systems
10 Henry Hermans (OUN), OpenU: design of an integrated system to support lifelong learning
11 Yongming Luo (TUE), Designing algorithms for big graph datasets: A study of computing bisimulation and joins
12 Julie M. Birkholz (VU), Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks
13 Giuseppe Procaccianti (VU), Energy-Efficient Software
14 Bart van Straalen (UT), A cognitive approach to modeling bad news conversations
15 Klaas Andries de Graaf (VU), Ontology-based Software Architecture Documentation
16 Changyun Wei (UT), Cognitive Coordination for Cooperative Multi-Robot Teamwork
17 André van Cleeff (UT), Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs
18 Holger Pirk (CWI), Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories
19 Bernardo Tabuenca (OUN), Ubiquitous Technology for Lifelong Learners
20 Lois Vanhée (UU), Using Culture and Values to Support Flexible Coordination
21 Sibren Fetter (OUN), Using Peer-Support to Expand and Stabilize Online Learning
22 Zhemin Zhu (UT), Co-occurrence Rate Networks
23 Luit Gazendam (VU), Cataloguer Support in Cultural Heritage
24 Richard Berendsen (UVA), Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation
25 Steven Woudenberg (UU), Bayesian Tools for Early Disease Detection
26 Alexander Hogenboom (EUR), Sentiment Analysis of Text Guided by Semantics and Structure
27 Sándor Héman (CWI), Updating compressed colomn stores
28 Janet Bagorogoza (TiU), Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO
29 Hendrik Baier (UM), Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains
30 Kiavash Bahreini (OU), Real-time Multimodal Emotion Recognition in E-Learning
31 Yakup Koç (TUD), On the robustness of Power Grids
32 Jerome Gard (UL), Corporate Venture Management in SMEs
33 Frederik Schadd (TUD), Ontology Mapping with Auxiliary Resources
34 Victor de Graaf (UT), Gesocial Recommender Systems
35 Jungxao Xu (TUD), Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction

---

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow

03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
04 Laurens Rietveld (VU), Publishing and Consuming Linked Data
05 Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UVA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VU), Refining Statistical Data on the Web
19 Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UVA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UVA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (UvT), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UVA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime

02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation

03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines

04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store

05 Mahdieh Shadi (UVA), Collaboration Behavior

06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search

07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly

08 Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery

09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text

10 Robby van Delden (UT), (Steering) Interactive Play Behavior

11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment

12 Sander Leemans (TUE), Robust Process Mining with Guarantees

13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology

14 Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior

15 Peter Berck (RUN), Memory-Based Text Correction

16 Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines

17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution

18 Ridho Reinanda (UVA), Entity Associations for Search

19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval

20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility

21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)

22 Sara Magliacane (VU), Logics for causal inference under uncertainty

23 David Graus (UVA), Entities of Interest — Discovery in Digital Traces

24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning

25 Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search

26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch

27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors

28 John Klein (VU), Architecture Practices for Complex Contexts

29 Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

30 Wilma Latuny (UvT), The Power of Facial Expressions

31 Ben Ruijl (UL), Advances in computational methods for QFT calculations

32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives

33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity

34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics

35 Martine de Vos (VU), Interpreting natural science spreadsheets

36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging

37 Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy

38 Alex Kayal (TUD), Normative Social Applications

39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems

41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle

42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets

43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval

44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering

45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement

46 Jan Schneider (OU), Sensor-based Learning Support

47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration

48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations

02 Felix Mannhardt (TUE), Multi-perspective Process Mining

03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction

04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks

05 Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process

06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems

07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems

08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems

09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations

10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology

11 Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks

12 Xixi Lu (TUE), Using behavioral context in process mining

13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future

14 Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters

15 Naser Davarzani (UM), Biomarker discovery in heart failure

16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children

17 Jianpeng Zhang (TUE), On Graph Sample Clustering

18 Henriette Nakad (UL), De Notaris en Private Rechtspraak

19 Minh Duc Pham (VUA), Emergent relational schemas for RDF

20 Manxia Liu (RUN), Time and Bayesian Networks

21 Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games

22 Eric Fernandes de Mello Araujo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks

23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis

24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots

25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections

26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology

27 Maikel Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis

28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel

29 Yu Gu (UVT), Emotion Recognition from Mandarin Speech

30 Wouter Beek, The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification

02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty

03 Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data Sources

04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data

05 Sebastiaan van Zelst (TUE), Process Mining with Streaming Data

06 Chris Dijkshoorn (VU), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms

08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes

09 Fahimeh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems

10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction

11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs

12 Jacqueline Heinerman (VU), Better Together

13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation

14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses

15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments

16 Guangming Li (TUE), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models

17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts

18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication

19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents

20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence

21 Cong Liu (TUE), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection

22 Martin van den Berg (VU),Improving IT Decisions with Enterprise Architecture

23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification

24 Anca Dumitrache (VU), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing

25 Emiel van Miltenburg (VU), Pragmatic factors in (automatic) image description

26 Prince Singh (UT), An Integration Platform for Synchromodal Transport

27 Alessandra Antonaci (OUN), The Gamification Design Process applied to (Massive) Open Online Courses

28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations

29 Daniel Formolo (VU), Using virtual agents for simulation and training of social skills in safety-critical circumstances

30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems

31 Milan Jelisavcic (VU), Alive and Kicking: Baby Steps in Robotics

32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games

33 Anil Yaman (TUE), Evolution of Biologically Inspired Learning in Artificial Neural Networks

34 Negar Ahmadi (TUE), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES

35 Lisa Facey-Shaw (OUN), Gamification with digital badges in learning programming

36 Kevin Ackermans (OUN), Designing Video-Enhanced Rubrics to Master Complex Skills

37 Jian Fang (TUD), Database Acceleration on FPGAs

38 Akos Kadar (OUN), Learning visually grounded and multilingual representations

2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour

02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models

03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding

04 Maarten van Gompel (RUN), Context as Linguistic Bridges

05 Yulong Pei (TUE), On local and global structure mining

06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support

07 Wim van der Vegt (OUN), Towards a software architecture for reusable game components

08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search

09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research

10 Alifah Syamsiyah (TUE), In-database Preprocessing for Process Mining

11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models

12 Ward van Breda (VU), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment

13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming

14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases

15 Konstantinos Georgiadis (OUN), Smart CAT: Machine Learning for Configurable Assessments in Serious Games

16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling

17 Daniele Di Mitri (OUN), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences

18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems

19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems

20 Albert Hankel (VU), Embedding Green ICT Maturity in Organisations

21 Karine da Silva Miras de Araujo (VU), Where is the robot?: Life as it could be

22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar

23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging

24 Lenin da Nobrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots

25 Xin Du (TUE), The Uncertainty in Exceptional Model Mining

26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization

27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context

28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality

29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference

30 Bob Zadok Blok (UL), Creatief, Creatieve, Creatiefst

31 Gongjin Lan (VU), Learning better – From Baby to Better

32 Jason Rhuggenaath (TUE), Revenue management in online markets: pricing and online advertising

33 Rick Gilsing (TUE), Supporting service-dominant business model evaluation in the context of business model innovation

34 Anna Bon (MU), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development

35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

---

2021 01 Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space

02 Rijk Mercuur (TUD), Simulating Human Routines:Integrating Social Practice Theory in Agent-Based Models

03 Seyyed Hadi Hashemi (UVA), Modeling Users Interacting with Smart Devices

04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning

05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems

06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot

07 Armel Lefebvre (UU), Research data management for open science

08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking

09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Childrenï¿œs Collaboration Through Play

10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning

11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision

12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs

13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning

14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support

15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm

16 Esam A. H. Ghaleb (UM), BIMODAL EMOTION RECOGNITION FROM AUDIO-VISUAL CUES

17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks

18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks

19 Roberto Verdecchia (VU), Architectural Technical Debt: Identification and Management

20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems

21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes

---