HealthInsights

An Online Conversational Survey for Understanding Worker Health in Crowdsourcing Platforms

Qiu, Sihang; Gadiraju, Ujwal; Zheng, Xiaolong

DOI
10.1145/3729176.3729201

Publication date
2025

Document Version
Final published version

Published in
CHIWORK 2025 - Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work

HealthInsights: An Online Conversational Survey for Understanding Worker Health in Crowdsourcing Platforms

Sihang Qiu*
Web Information Systems, EEMCS
Delft University of Technology
Delft, Netherlands
sihangq@acm.org

Ujwal Gadiraju
Web Information Systems, EEMCS
Delft University of Technology
Delft, Netherlands
u.k.gadiraju@tudelft.nl

Xiaolong Zheng
The School of Artificial Intelligence
University of Chinese Academy of Sciences
Beijing, China
The State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation
Chinese Academy of Sciences
Beijing, China
xiaolong.zheng@ia.ac.cn

Abstract

Crowdsourcing marketplaces have gradually flourished over the last decade. With the growing landscape of online work in general, and the rise of paid microtask crowdsourcing in particular, the health and wellbeing of crowd workers has become an important concern. In this paper, we present an online conversational survey, named `HealthInsights`, for understanding the status quo of workers' health-related background, physical health, mental health, and their needs. We carried out a study on two popular platforms – Mechanical Turk and Prolific. Results show that the survey has acceptable reliability and validity. 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. Our work has important implications on task and workflow design that are centered around worker health on crowdsourcing platforms.

CCS Concepts

• **Information systems** → **Crowdsourcing**.

Keywords

Crowdsourcing, Microtask, Health, Wellbeing, Interventions

ACM Reference Format:
Sihang Qiu, Ujwal Gadiraju, and Xiaolong Zheng. 2025. HealthInsights: An Online Conversational Survey for Understanding Worker Health in Crowdsourcing Platforms. In *CHIWORK '25: Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work (CHIWORK '25), June 23–25, 2025, Amsterdam, Netherlands.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3729176.3729201

*Sihang Qiu is currently affiliated with the State Key Laboratory of Digital-Intelligent Modeling and Simulation and NUDT, China. This work was primarily done while Sihang Qiu was at the Delft University of Technology.

CHIWORK '25, Amsterdam, Netherlands

ACM ISBN 979-8-4007-1384-2/25/06
https://doi.org/10.1145/3729176.3729201

1 Introduction

Over the years, researchers and practitioners have spent a significant amount of effort in improving quality-related outcomes in microtask crowdsourcing. From adaptive task assignment [43, 88], to worker modeling and prescreening methods [35, 74, 89, 90], to response aggregation [12, 70, 73] and task pricing [15, 25], and creative incentive schemes [31, 53, 58], many methods have been proposed to optimize quality related outcomes and minimize costs.

Owing to the central role that microtask crowdsourcing plays in research and industry, recent studies have also 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 creating awareness about the invisible labor that prevails on microtask crowdsourcing platforms [41]. Prior works have proposed tools to help crowd workers address issues related to power asymmetry and worker invisibility on platforms [47, 48, 96]. Researchers have highlighted the impact of work rejection and the importance of facilitating trust between workers and requesters [68, 93]. Others have built platforms to facilitate collective action [92], and proposed methods to help crowd workers share their risks alongside rewards on crowdsourcing platforms [30]. Recent work has proposed methods to automatically ensure fair pay for crowd workers [115], and reduce the negative impact of exposure to harmful content [23]. People are already aware of the fact that worker health is in general neglected in the crowdsourcing ecosystem, however, little is currently understood about the status quo of workers' health and how this varies across crowdsourcing platforms. Creating a qualitative and quantifiable understanding of worker health across platforms is an important first step toward addressing existing issues. However, 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.

Furthermore, 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 "microtasks" that they are asked to complete, with an aim to maintain and improve their health, and increase 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 to understand what workers prefer in terms of the intervention types, their duration, and frequency.

Therefore, we aim to address the following research questions:

***RQ1**: How can the prevalent physical and mental health status of crowd workers be acquired through a conversational survey?*

***RQ2**: To what extent are healthcare interventions needed on crowdsourcing platforms?*

To this end, our research objective is to design a conversational survey to comprehensively investigate workers' physical and mental health, as well as their needs for healthcare interventions. The main focus is to validate the reliability and validity of the designed conversational survey, and then to conduct a preliminary analysis on survey results. Therefore, we designed `HealthInsights` – 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 a conversational user interface to gather survey responses from crowd workers, as prior studies have shown that conversational user interfaces are effective alternatives to traditional web-based surveys, as they can increase satisfaction and engagement, while improving the overall survey quality [19, 22, 50, 61, 82]. We conducted the study on two popular crowdsourcing platforms – Amazon Mechanical Turk (MTurk) and Prolific. Results show that the survey has acceptable reliability and validity, and therefore, effectively reflects workers' health status and needs on the two platforms. We found 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. Workers on MTurk in general reported significantly better physical health status compared to workers on Prolific. In terms of their mental health, we found that, surprisingly, workers on MTurk reported significantly worse mental health status compared to workers on Prolific.

Our findings suggest that, for a better future of workers in crowdsourcing marketplaces, task requesters should design and publish tasks in a healthy and platform-specific manner (e.g., pushing short physical exercises for Prolific workers; pushing breaks and relaxing interventions for MTurk workers). We also 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.

2 Related Work

2.1 Health and 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 [105]. The report emphasized the important role that wellbeing can play [9]. A recent study systematically reviewed the relationship between the office working environment and employee health & wellbeing [20]. 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 [33]. Furthermore, recent work has shown that the state-of-the-art methods can to some extent improve work productivity and wellbeing [116]. Another direction of relevant research relates to worker emotions and moods [4, 24, 108], since both emotion and mood are valenced affective responses. Prior studies have proposed a variety of instruments to measure the emotion, such as the Self Assessment Manikin [7], the Affective Slider [5], the Achievement Emotions Questionnaire [77], and Pick-a-Mood (PAM) [24]. Based on the worker moods measured by PAM, researchers have presented 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 [34, 83, 119, 123]. In this work, we design `HealthInsights` – an online survey – to deeply investigate workers' health and wellbeing.

2.2 Physical Discomforts Experienced by Workers

Physical discomforts and ergonomics of office work, particularly for sitting workers, have become an important research topic for decades [72]. Researchers have started to invent techniques and instruments to effectively assess working postures, and posture-related somatic problems [21, 40]. Recent studies started to focus on ergonomics and working postures while using computers [62, 85, 117]. In prior work by Luttmann et al., the authors performed precise measurements to assess muscular activities and working conditions [62]. 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 [85]. Woo et al. performed a systematic review to propose ergonomics standards and guidelines for computer workstation design, and summarize their effect on worker health [117]. Since crowdsourcing is a common computer-based work mode, it is important to consider how this type of labor impacts the health of its workers. In the context of crowdsourcing, researchers have found that both physical and mental fatigue could have negative impacts on crowd work [64, 122]. 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.

2.3 Mental Health of Workers

Mental health has become a very important topic in society since it relates to everyone in the world [75]. There are many previous studies focusing on mental disorders and corresponding treatments (e.g. meditation, hypnosis, relaxation therapies, etc.) relating to office work [59, 71, 100, 104, 110]. To assess one's overall health and mental health, the SF-36 survey has been extensively used [112, 113], 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) [55]. In addition to the prevailing platform dynamics in crowdsourcing marketplaces [92], 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 [13, 39, 78, 98]. 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 [87, 99]. Research has shown that content moderators are regularly exposed to far more malignant content [14]. To address this problem, researchers have used blurring to reduce the harmful content exposure time for moderators [23]. In this work, we will assess workers' mental health and psychosocial working environment and make comparisons across crowdsourcing platforms, to better inform health-related interventions for the future.

2.4 Power Asymmetry in Crowdsourcing Marketplaces

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 building a healthy requester-worker relationship [33, 41, 67, 93]. Another popular research topic relates to improving workers' possibilities of development [6, 16, 101]. Atelier was therefore designed to re-purpose crowdsourcing tasks as mentored and paid skill development, named micro-internships [101]. Chiang et al. designed a system called Crowd Coach to assist workers in skill growth [16]. Others have proposed a variety of mechanisms to improve trust and ensure fair payment in crowdsourcing marketplaces [37, 42, 91, 94, 115]. Whiting et al. developed a tool for task requester to ensure the minimum wage [115]. Recent work by Savage et al. recommended transparency criteria to guide workers to earn higher salaries [94]. Furthermore, prior works attempted to improve worker wellbeing with a better task design, by improving task clarity [36, 49, 63, 118], and combining workers' opinions [8, 97]. Therefore, we use `HealthInsights` to not only understand how healthy crowd workers are but also listen to what they really need to improve the working environment.

2.5 Conversational Crowdsourcing

Conversational crowdsourcing employs conversational agents to help crowd workers complete tasks, which has become increasingly important, particularly after large language models emerged [2]. Prior works designed multiple conversational agents leveraging inputs of human users recruited from crowdsourcing marketplaces, to address general knowledge tasks [44, 45, 57]. Conversational crowdsourcing was proposed as a novel paradigm recently, which facilitates the execution of different types of popular crowdsourcing tasks, showing the effectiveness of using conversational agents or interfaces in terms of improving user satisfaction, user engagement, output quality, user moods, and many other subjective worker perceptions [50, 81–83]. Apart from crowdsourcing marketplaces, conversational crowdsourcing can also be deployed on social network platforms, combined with messaging applications, such as Facebook and Twitter [95, 106]. In this work, based on the important findings discovered by previous studies, we designed a conversational agent enabling survey completion on typical crowdsourcing platforms.

3 Survey Design

We designed `HealthInsights` – a survey consisting of 60 items, delivered through the 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. Furthermore, we employed a conversational user interface to gather survey responses.

3.1 Worker Background

In the first part of the survey, we ask background questions to understand the demographic information and working environment of crowd workers. As listed in Table 1, 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. Next, we aimed to understand workers' general working environment and immediate working environment [72] using questions 8-11. 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 [10]. Finally, we asked workers if they take any measures to keep themselves healthy.

3.2 Working Ergonomics and Physical Health

The second part of our survey addresses workers' working ergonomics and physical health. Based on the prior studies on ergonomics [40, 62, 117] and Stanford's computer workstation ergonomics self-evaluation form [28], we designed a survey that covers the relevant aspects of working ergonomics — chair, keyboard and mouse, screen/monitor, breaks/practices, and overall posture (question 15-23 in Table 2). Questions in all aspects were selected from validated surveys used in previous studies [28, 40, 62, 117].

We first asked workers about their working postures (sitting, standing, or other postures). Using example pictures of healthy

Table 1: The questions used in the first part of the survey: 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

Table 2: The questions used in the second part of the survey: 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?	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

sitting/standing postures, we asked workers to rate the degree of health of their 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, 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

Table 3: The questions used in the third part of the survey: 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 MTurk/Prolific or other worker forums)?	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

position of the screen top. As prescribed by recent work [117], 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. According to the categories of different body parts used in ergonomics research [40], to measure physical health, we asked workers to rate their perceived degree of comfort (on an average working day) with respect to each body part (questions 24-29). 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.

3.3 Psychosocial Condition and Mental Health

We investigate 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. [38, 54]. In this study, we used subsets of existing instruments to measure pyschosocial conditions, mental wellbeing, and working energy/fatigue [55, 112, 113] as shown in Table 3.

First, we selected 10 representative questions from the COPSOQ CORE [55, 112, 113] 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 3). 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 [52]. *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 [55]. To measure worker mental health, in total 10 questions were selected from SF-36 [112, 113]. One question is for self-reporting general health. 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.

3.4 Workers' Needs

The last part of the survey is devoted to the inquiry 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

Table 4: The questions used in the fourth part of the survey: 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

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 4.

The first two questions are about workers' needs with regard to physical health and mental health respectively. 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, preferred working mode, 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, the size of the intervention (length in minutes), the frequency of the interventions, and whom they believe this responsibility lies with.

4 Conversational Interface Implementation

Conversational user interfaces have become increasingly common as an alternative to the traditional graphical user interfaces. Research has shown that a conversational interface or a chatbot is capable of providing better user engagement and satisfaction [19, 22, 61, 82] and has the potential of leading to a better output quality [50], due to its human-like means of interaction. Conversational interfaces have been successfully employed in a variety of domains [120], ranging from design [107] to search [3, 51, 111]. 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 [56, 69, 86], and for mental-health support or treatment [1, 32, 60, 79, 102]. Famous chatbots include ELIZA, Woebot, etc.[80, 114].

In this study, we employed a conversational user interface to guide workers to complete the survey, instead of using traditional web-based survey forms. We implemented the online conversational survey using our web-based conversational survey tool[1] called TickTalkTurk [84]. The tool provides a GUI where one can directly input questions, and then generate a conversational web-based user interface for survey execution. A screenshot of the conversational interface for completing the survey is shown in Figure 1. Conversational elements such as greetings, the response delay, and repeating the worker response are applied on the chatbot to improve conversational experience [103]. On successfully completing the survey, the conversational agent provides workers with a survey completion code, which they can use to earn their rewards.

5 Experimental Setup

In this study, we focus on English-based platforms to survey crowd workers and make comparisons on their health status. Among the three mainstream English platforms (MTurk, Prolific, and Appen) [76], we chose two platforms – MTurk and Prolific, due to

[1] https://github.com/qiusihang/ticktalkturk

Figure 1: A screenshot of the conversational interface employed to gather survey responses from workers.

their different usage profiles. MTurk is commonly used for large-batch data labeling tasks [26], while Prolific is commonly used for user studies and market research. We do not consider Appen in this study since it features similar types of data labeling tasks (compared with MTurk), and requires enterprise-level subscriptions. 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. This study has been approved by the human research ethics committee of our institute.

5.1 Quality Control

Our broad goal is to see the reliability and validity of `HealthInsights`, and to understand the health of general crowd workers on different crowdsourcing platforms. 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 [66]. 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 (which was not indicated before they started the survey).

5.2 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 around 14.2 minutes (850.28 ± 324.35 seconds). That is $\tilde{1}6.0$ minutes (961.43 ± 334.29 seconds) on Prolific and $\tilde{1}2.1$ minutes (728.43 ± 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.

6 Results

All the workers completing the survey will be given a completion code by default. We applied the same working mode on MTurk and Prolific. 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. All of the results and the code of `HealthInsights` are publicly available[2].

[2]https://sites.google.com/view/workerhealth/

6.1 Demographics

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 2 shows the workers' country of residence. Prolific users came from Europe, North America, South America, Africa, and Asia. The majority of the workers (94) came from Europe (35 from the United Kingdom; 31 from Portugal; 15 from Poland; and other 13 European countries). On MTurk, 93 out of 135 workers reported to be residing in the United States. As shown in Figure 2, the workers on Prolific were generally younger and less experienced compared to the MTurk workers. (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). Note that in the survey we asked about worker experience on either Prolific or MTurk, rather than asking experience of general crowdsourcing. The reason is that we do have concerns that many workers on Prolific may not precisely understand the term "crowdsourcing", since Prolific advertises itself as an online survey platform. 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).

Investigating workers' usual working time is an important part of understanding worker health. As shown in Figure 3, 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 3. 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. Furthermore, to have a better understanding of workers' working environment, we asked whether workers felt comfortable and healthy about their working environment (lighting, temperature, humidity, noise, etc.) and working devices (devices and setups in terms of control, display, compatibility, layout, posture, etc.) [72]. As shown in Figure 4, Prolific workers reported worse working environments 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, Shapiro-Wilk tests). In addition, we found that on both platforms self-reported health scores (Prolific: 4.84 ± 1.36, MTurk: 5.39 ± 1.26) are lower than self-reported comfort scores (Prolific: 5.26 ± 1.40, MTurk: 5.54 ± 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 both crowdsourcing platforms.

6.2 Survey Reliability and Validity

Although the `HealthInsights` survey is constructed by multiple verified questionnaires, the combination of their subsets, particularly presented in a conversational way, might affect the reliability and validity on crowdsourcing platforms [17]. Therefore, in this work, we investigate survey reliability and validity, on the parts for measuring worker health (questions 24-29 and 41-49). We do not carry out reliability and validity analysis on the parts of background, worker needs, and other exploratory items. Future work will further optimize the survey design.

We used Cronbach's alpha to measure the reliability of the survey. For all the items measuring mental wellbeing, the Cronbach's $\alpha = 0.784$ (95% confidence interval: 0.742 – 0.822). For all the items measuring fatigue and energy levels, the Cronbach's $\alpha = 0.686$ (95% CI: 0.621 – 0.741). In terms of physical health, the Cronbach's $\alpha = 0.862$ (95% CI: 0.836 – 0.886). The survey in general shows acceptable reliability.

We also performed factor analysis to investigate the survey validity of `HealthInsights` (also for the parts for measuring worker health). First, the Kaiser-Meyer-Olkin (KMO) test was applied to measure the sampling adequacy. The KMO value was 0.844, showing that the preliminary results were good for conducting factor analysis. Afterward, the Bartlett's test of Sphericity was carried out to see whether the correlation matrix is an identity matrix. The results showed that $p = 0.000$ and $\chi^2 = 1954.72$. It means that the correlation matrix isn't an identity matrix, representing there are correlations among survey items and it is meaningful to do factor analysis.

The results of exploratory factor analysis (EFA) are shown in Table 5. There are no specific rules in terms of factor selection and prior studies usually chose factors whose eigenvalues are greater than 1, or when their cumulative percentage of variance reaches a specific threshold [29]. In Table 5, we could see that there are three factors having eigenvalues obviously larger than all the others. Furthermore, in this study, we intended to use `HealthInsights`

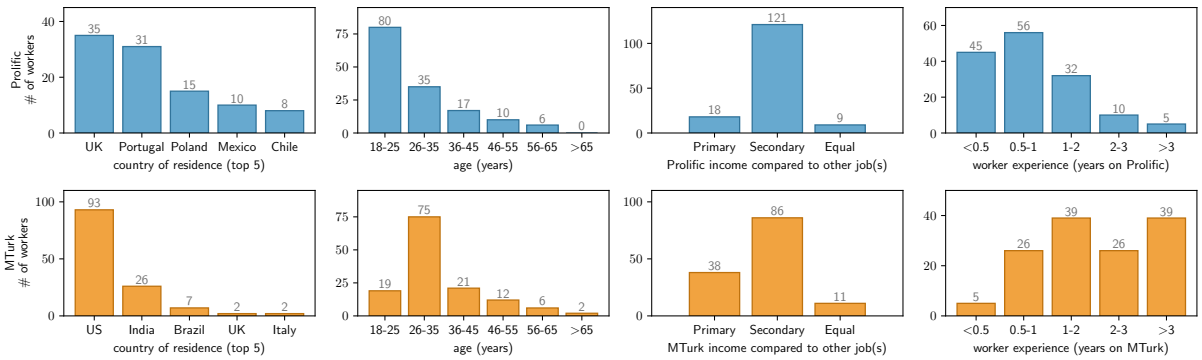

Figure 2: Worker demographics in terms of the country of residence, age, income, and working time on two crowdsourcing platforms (Prolific and MTurk).

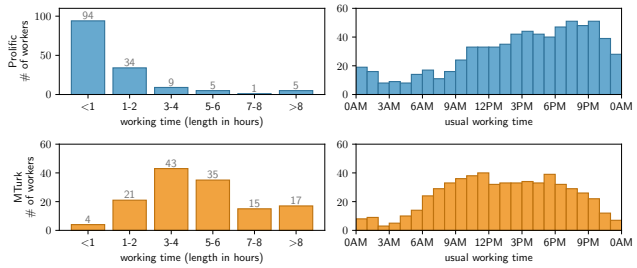

Figure 3: Workers' usual working time throughout a day on two crowdsourcing platforms (Prolific and MTurk).

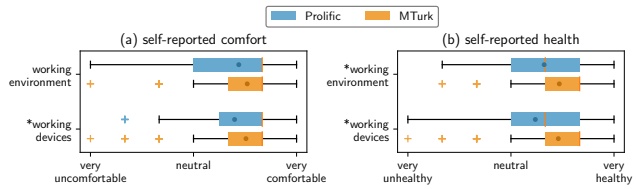

Figure 4: Self-reported (a) comfort and (b) health in terms of working environment (lighting, temperature, humidity, noise, etc.) and working devices (control, display, compatibility, layout, posture, etc.) on two crowdsourcing platforms (Prolific and MTurk), showing MTurk workers in general report significantly better (*p<0.05) working environments and devices in comparison to Prolific workers.

to measure three dimensions of worker health – mental wellbeing, fatigue/energy levels, and physical discomforts. Therefore, we selected three factors and applied orthogonal rotation to see item loadings (Table 6). We can clearly see how each item loads on three factors. The items for measuring physical health apparently have higher loadings (> 0.62) on factor 2, while items for measuring mental health (wellbeing and fatigue) have higher loadings on factor 1 and factor 3. Factors 1&3 together represent mental health. However, they do not clearly separate wellbeing-related and fatigue-related items, implying mental health is a rather complex concept to understand. Future research could improve the subset selected from SF-36 (where the survey items originally came from) to precisely measure mental health in various aspects. Factor 2 undoubtedly represents physical health, showing that `HealthInsights` has good validity in terms of explaining worker health from physical and mental perspectives, respectively.

Table 5: Top 10 factors' eigenvalues, percentages of variance, and cumulative percentages of variances.

Factor	Eigenvalue	% of variance	Cumulative %
1	5.34	35.6%	35.6%
2	2.38	15.9%	51.5%
3	1.69	11.3%	62.8%
4	0.800	5.3%	68.1%
5	0.710	4.7%	72.8%
6	0.635	4.2%	77.1%
7	0.591	3.9%	81.0%
8	0.543	3.6%	84.6%
9	0.442	2.9%	87.6%
10	0.392	2.6%	90.2%

6.3 An Overview of Survey Results

This work presents an overview of survey results investigating the feasibility of `HealthInsights` for measuring worker health on crowdsourcing platforms. While we explore the survey data, we do not delve deeply into all exploratory items. Future research could leverage this survey in large-scale online crowdsourcing experiments, utilizing all items for a comprehensive causal analysis to identify factors contributing to unhealthy situations.

6.3.1 Ergonomics and Physical Health. A proper working posture is essential to one's health. As shown in Figure 5 (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 to workers and asked them to rate the degree of health of their overall working postures by comparing

Table 6: The loadings of survey items on the three selected factors.

Survey item	Factor 1	Factor 2	Factor 2
Mental Health – Wellbeing 1	0.720	-0.150	-0.003
Mental Health – Wellbeing 2	0.836	0.006	-0.152
Mental Health – Wellbeing 3	-0.372	0.085	0.698
Mental Health – Wellbeing 4	0.818	-0.091	-0.184
Mental Health – Wellbeing 5	-0.212	0.107	0.840
Mental Health – Fatigeu 1	-0.051	0.146	0.778
Mental Health – Fatigeu 2	-0.092	0.135	0.800
Mental Health – Fatigeu 3	0.754	-0.246	-0.165
Mental Health – Fatigeu 4	0.620	-0.237	-0.249
Physical Health – Eyes	-0.131	0.764	0.054
Physical Health – Head	-0.192	0.764	0.110
Physical Health – Neck/shoulders	-0.145	0.819	0.015
Physical Health – Back	-0.042	0.803	0.134
Physical Health – Seat/thighs	-0.076	0.738	0.129
Physical Health – Knees/feet	-0.130	0.620	0.155

the examples with their own working postures. Results revealed that, as shown in Figure 5 (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 (56 healthy vs 64 unhealthy) 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 (c). As for working ergonomics, we 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 6 (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 [117]. Moreover, we found that many workers (42 on Prolific and 47 on MTurk) reported their screen tops were higher than the eye level (Figure 6 (b)). This is possibly due to the large screen size. With regard to the keyboard and the mouse (Figure 6 (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 [117].

Results of physical health are shown in Figure 7. Significance testings suggested that for both Prolific and MTurk ($H = 70.35, p < 0.001$ and $H = 20.49, p = 0.001$ respectively, Kruskal-Wallis H-test), median comfort scores of different body parts were not equal. As we can see from Figure 7, 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.

6.3.2 Psychosocial Condition and Mental Health. In the third part of the survey, we focus on the psychosocial condition and mental health. Results are reported in Table 7. The score ranges from 1 to 5. A higher score represents a better psychosocial condition. 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 show 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).

Table 7: Psychosocial scores of workers on Prolific and MTurk platforms, showing significant differences (*p<0.05) in 7 out of 10 dimensions, particularly in work pace, emotional demands, social support from workers, and feedback at work.

Dimension	Prolific	MTurk
Quantitative demands	3.82 ± 0.85	3.90 ± 0.75
*Work pace (R)**	3.58 ± 0.99	2.89 ± 1.14
*Emotional demands (R)**	4.22 ± 0.89	3.24 ± 1.23
Influence at work	3.38 ± 1.28	3.44 ± 1.12
Possibilities for development	3.42 ± 1.17	3.59 ± 1.10
*Meaning of work**	3.78 ± 1.05	3.48 ± 1.18
*Social support from requesters**	2.69 ± 1.30	3.02 ± 1.18
*Social support from workers**	1.96 ± 1.21	2.97 ± 1.19
*Feedback at work**	2.44 ± 0.96	3.30 ± 0.89
*Sense of community**	3.15 ± 1.02	3.44 ± 1.02

The health scores of mental wellbeing and energy/fatigue are displayed as boxplots in Figure 8. The health score ranges from 0 to 100. A higher score represents a healthier mental condition.

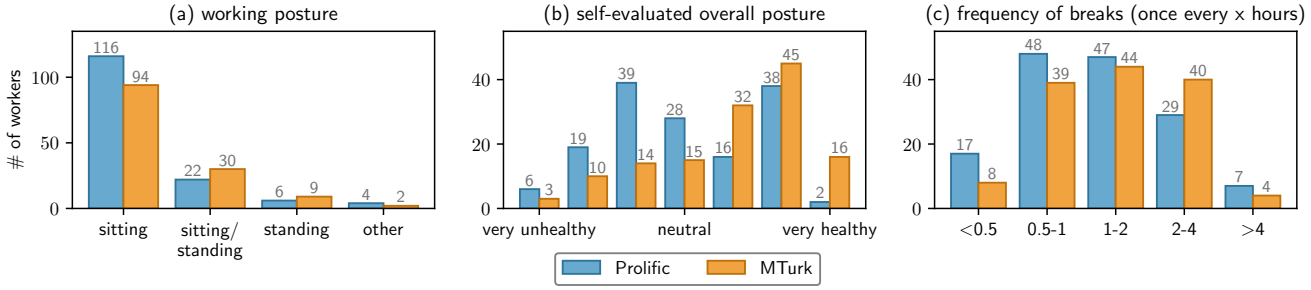

Figure 5: Self-evaluated working postures and physical health across two crowdsourcing platforms (Prolific and MTurk).

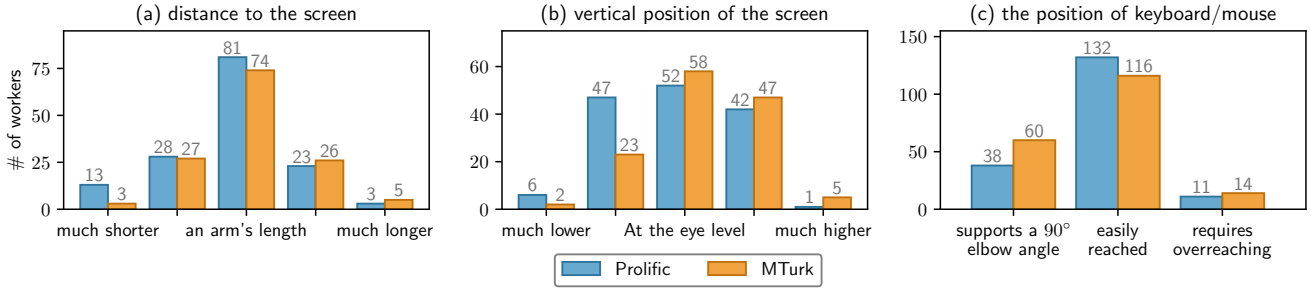

Figure 6: Self-evaluated positions of working devices across two crowdsourcing platforms (Prolific and MTurk).

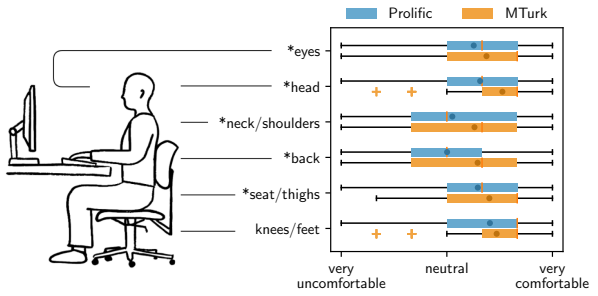

Figure 7: Physical health scores across two crowdsourcing platforms (Prolific and MTurk), showing MTurk workers report significantly better comfort (*p<0.05) in five of six body areas compared to Prolific workers, particularly in neck-/shoulders and back regions.

Interestingly, in terms of mental health, we found completely opposite results compared to physical health (Figure 7). 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. In comparison with baseline values provided by RAND (fatigue: 52.15 ± 22.39 and mental wellbeing: 70.38 ± 21.97), we found that MTurk workers have worse mental wellbeing conditions in comparison with the average.

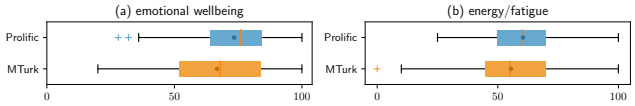

Figure 8: Health scores of workers' mental wellbeing and energy/fatigue (subsets of SF-36) across two crowdsourcing platforms (Prolific and MTurk).

6.4 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 9 (a). In terms of needs about mental health, as we can see from Figure 9 (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.

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 whether workers would be happy to receive the interventions. As displayed in Figure 10, 82 workers (61%) on MTurk and 81 workers (55%) on Prolific, respectively, reported that they would be happy to receive 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.

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 providing interventions to improve their health. Results revealed that 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, a large proportion (63%) of workers on both platforms reported that they were okay with not getting paid. Furthermore, the length of intervention minutes should not be neither too long (> 10 minutes) nor too short (< 1 minute) according to the results. 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 (72% on Prolific and 62% on MTurk) "agreed" that the crowdsourcing platforms should be responsible for sending interventions to improve their health.

7 Discussion

7.1 A Guide to Using `HealthInsights`

Analysis of reliability and validity showed that `HealthInsights` could be effectively published as microtasks on crowdsourcing platforms to consistently collect high-quality user responses. Furthermore, the whole survey has plenty of questions to acquire worker background information, ergonomic working environments, and psychosocial conditions. These data can be used to conduct correlation analysis or causal analysis among different items, thus helping us understand worker health better. In addition, we emphasize the importance of `HealthInsights` with regard to its goal of understanding worker needs. It is necessary to know what kinds of interventions are really needed, and how should we provide such health-related interventions to crowd workers. Each part of `HealthInsights` can be used separately to understand a specific dimension of worker health.

Furthermore, it is worth mentioning that we found that some workers were fond of completing the online survey with a conversational interface (although we did not ask for their opinions about it), since it made them feel 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.

7.2 General Health Status of Workers

Workers from MTurk and Prolific reported similar health-related issues. In terms of physical health, crowd workers from both platforms felt less comfortable in their necks, shoulders, and backs. MTurk workers in our study reported a relatively better physical health status but relatively worse mental health status compared to Prolific workers. Furthermore, tasks on MTurk were reported to be more emotionally demanding, and required a faster working pace, and tasks on Prolific were reported to be more meaningful. However, MTurk workers received more social supports/feedback. Results pertaining to the health analysis revealed that workers across both platforms, share a common pattern of health status.

7.3 Implications

Herbert Marcuse's concept of one-dimensional discourse could explain how crowdsourcing platforms and practitioners emphasize technical advantages such as flexibility and autonomy to obscure potential health risks, making worker suffering invisible [65]. This calls for systemic change pushed by crowdsourcing platforms and requesters, rather than workers' individual adaptation.

7.3.1 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 or legally protected. As more and more people turn towards crowdsourcing marketplaces (and the broader spectrum of online work) to support their livelihood, crowdsourcing platforms 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

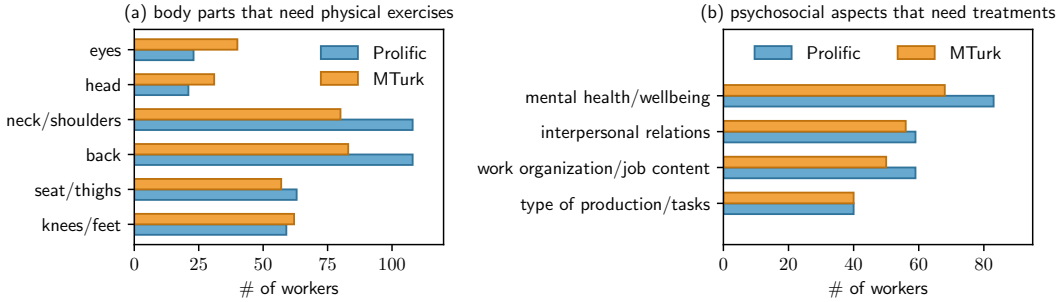

Figure 9: Worker needs in terms of (a) physical health and (b) mental health respectively across two crowdsourcing platforms (Prolific and MTurk).

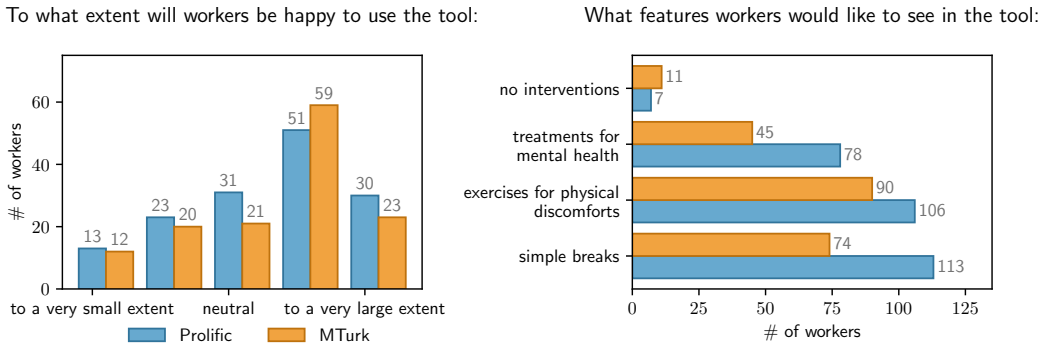

Figure 10: Worker needs with regard to a health-related tool that can provide interventions (breaks/exercises/treatments) across two crowdsourcing platforms (Prolific and MTurk).

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 questions need to be addressed before such a reality can be realized. How can health interventions be introduced and packaged between or within microtasks? To what extent would such interventions serve as an effective means to improve worker health and wellbeing as a result of crowd work?

7.3.2 For task requesters. According to the results of mental health questions, the task content also plays an important role in worker health. Experiencing meaningfulness is a critical psychological state derived from the job characteristics model [46]. A simple starting point for task requesters, especially for requesters who publish tasks on MTurk, is to emphasize the meaning of the task and slow down the working pace, rather than giving them pressure and letting them work like robots, as suggested by Chandler et al. [11] as well. Furthermore, task requesters should consider involving more learning elements and reducing emotional demands (particularly for content moderation tasks) in crowd work, which has been proved to be effective in terms of improving their performance and mental wellbeing [23, 27, 99]. Task requesters can begin by shouldering some of the responsibility to ensure worker wellbeing. We envision that requesters can, for instance, provide MTurk workers with small health interventions for breaks and relaxation, and provide Prolific workers with some physical exercises. Such interventions can be designed as "tasks" to be completed and packaged together with a task batch. Paying crowd workers to consume such health interventions would result in increasing the costs for the task requesters by a relatively small fraction. We suggest that requesters should pay crowd workers to consume such health interventions although a large proportion of workers are fine with not getting paid. In return, task requesters can reap the benefits of having a healthy and sustainable workforce to rely on and mutually flourish.

7.4 Limitations and Future Work

In this study, we recruited 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 and Toloka), make comparisons with existing studies using the same questionnaires (SF-36, COPSOQ, etc.), or consider performing studies on online freelancing marketplaces, to make broader implications with regard to the entire online gig economy.

Reliability has been an important issue in crowdsourcing-related research [17]. Particularly, with the rapid development of LLM, tackling AI-generated content has become a major challenge [18, 109, 121], which may also affect the validity of `HealthInsight`. However, the health-related survey is different from conventional microtasking. In the future, completion of `HealthInsight` or other similar surveys should not be incentivized by monetary reward – rather, they should be designed in a way to motivate crowd workers to complete such surveys with an aim to improve their health. In such designs, LLM can play as personal assistants instead of cheating tools.

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.

While conversational interfaces offer an engaging way to collect survey data, various conversational styles, linguistic elements, and particularly AI-generated content (if in the future LLMs are used) may introduce biases and bring privacy issues. Future studies could address these limitations by exploring human-in-the-loop approaches to enhance the security, reliability, and generalizability of survey data collected via conversational interfaces.

8 Conclusion

In this work, we designed `HealthInsights` – a conversational survey, which consists of 60 items, to investigate and compare worker health across crowdsourcing platforms. We carried out experiments on Prolific and MTurk. First, we acquired data about workers' health-related background, physical health status, mental health status, and worker needs. Then, we performed analysis on the reliability and validity of `HealthInsights`. We found that MTurk workers reported significantly better physical health and working environment/devices, while Prolific workers reported significantly better mental health and wellbeing. Furthermore, according to the answers we acquired by explicitly asking workers about their needs, we proposed suggestions to crowdsourcing platforms and task requesters for improving worker health. This work has important implications in terms of facilitating a better and healthier working environment for future crowdsourcing.

Acknowledgments

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. This work was partially supported by the National Natural Science Foundation of China (nos. 62202477, 72225011, and 72434005).

References

[1] Tahir Abbas, Vassilis-Javed Khan, Ujwal Gadiraju, and Panos Markopoulos. 2020. Trainbot: A Conversational Interface to Train Crowd Workers for Delivering On-Demand Therapy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 3–12.

[2] Garrett Allen, Gaole He, and Ujwal Gadiraju. 2023. Power-up! what can generative models do for human computation workflows? *arXiv preprint arXiv:2307.02243* (2023).

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

[4] Christopher Beedie, Peter Terry, and Andrew Lane. 2005. Distinctions between emotion and mood. *Cognition & Emotion* 19, 6 (2005), 847–878.

[5] Alberto Betella and Paul FMJ Verschure. 2016. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PloS one* 11, 2 (2016).

[6] Jeffrey P Bigham, Kristin Williams, Nila Banerjee, and John Zimmerman. 2017. Scopist: building a skill ladder into crowd transcription. In *Proceedings of the 14th Web for All Conference on The Future of Accessible Work*. 1–10.

[7] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.

[8] Jonathan Bragg and Daniel S Weld. 2018. Sprout: Crowd-powered task design for crowdsourcing. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 165–176.

[9] Jenna Butler, Mary Czerwinski, Shamsi Iqbal, Sonia Jaffe, Kate Nowak, Emily Peloquin, and Longqi Yang. 2021. *Personal Productivity and Well-being–Chapter 2 of the 2021 New Future of Work Report*. Technical Report. Microsoft. https://aka.ms/newfutureofwork

[10] 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. 2002. Loneliness and health: Potential mechanisms. *Psychosomatic medicine* 64, 3 (2002), 407–417.

[11] Dana Chandler and Adam Kapelner. 2013. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization* 90 (2013), 123–133.

[12] Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2017. Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 5.

[13] Adrian Chen. 2012. Inside Facebook's Outsourced Anti-Porn and Gore Brigade Where Camel Toes are More Offensive Than Crushed Heads. *Gawker. Com* 16 (2012).

[14] Adrian Chen. 2014. The laborers who keep dick pics and beheadings out of your Facebook feed. *Wired* 23 (2014), 14.

[15] Justin Cheng, Jaime Teevan, and Michael S Bernstein. 2015. Measuring crowdsourcing effort with error-time curves. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1365–1374.

[16] Chun-Wei Chiang, Anna Kasunic, and Saiph Savage. 2018. Crowd coach: Peer coaching for crowd workers' skill growth. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–17.

[17] Michael Chmielewski and Sarah C Kucker. 2020. An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science* 11, 4 (2020), 464–473.

[18] Evgenia Christoforou, Gianluca Demartini, and Jahna Otterbacher. 2024. Generative AI in Crowdwork for Web and Social Media Research: A Survey of Workers at Three Platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 2097–2103.

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

[20] Susanne Colenberg, Tuuli Jylhä, and Monique Arkesteijn. 2020. The relationship between interior office space and employee health and well-being–a literature review. *Building Research & Information* (2020), 1–15.

[21] E Nigel Corlett and RP Bishop. 1976. A technique for assessing postural discomfort. *Ergonomics* 19, 2 (1976), 175–182.

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

[23] Anubrata Das, Brandon Dang, and Matthew Lease. 2020. Fast, Accurate, and Healthier: Interactive Blurring Helps Moderators Reduce Exposure to Harmful Content. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 33–42.

[24] Pieter MA Desmet, Martijn H Vastenburg, and Natalia Romero. 2016. Mood measurement with Pick-A-Mood: review of current methods and design of a pictorial self-report scale. *Journal of Design Research* 14, 3 (2016), 241–279.

[25] Djellel Difallah, Michele Catasta, Gianluca Demartini, and Philippe Cudré-Mauroux. 2014. Scaling-up the crowd: Micro-task pricing schemes for worker retention and latency improvement. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 2.

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

[27] Mira Dontcheva, Robert R Morris, Joel R Brandt, and Elizabeth M Gerber. 2014. Combining crowdsourcing and learning to improve engagement and performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3379–3388.

[28] Stanford University Ergonomics. [n. d.]. Stanford'scomputer workstation ergonomics self-evaluation form. https://ehs.stanford.edu/forms-tools/computer-workstation-ergonomics-evaluation

[29] Leandre R Fabrigar, Duane T Wegener, Robert C MacCallum, and Erin J Strahan. 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods* 4, 3 (1999), 272.

[30] Shaoyang Fan, Ujwal Gadiraju, Alessandro Checco, and Gianluca Demartini. 2020. CrowdCO-OP: Sharing Risks and Rewards in Crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–24.

[31] Oluwaseyi Feyisetan, Elena Simperl, Max Van Kleek, and Nigel Shadbolt. 2015. Improving paid microtasks through gamification and adaptive furtherance incentives. In *Proceedings of the 24th international conference on world wide web*. 333–343.

[32] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. 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 (2017), e19.

[33] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. 2017. 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 (2017), 1–29.

[34] Ujwal Gadiraju and Gianluca Demartini. 2019. Understanding Worker Moods and Reactions to Rejection in Crowdsourcing. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*. 211–220.

[35] Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. 2017. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 4 (2017), 1–26.

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

[37] Snehal Gaikwad, Durim Morina, Rohit Nistala, Megha Agarwal, Alison Cossette, Radhika Bhanu, Saiph Savage, Vishwajeet Narwal, Karan Rajpal, Jeff Regino, et al. 2015. Daemo: A self-governed crowdsourcing marketplace. In *Adjunct proceedings of the 28th annual ACM symposium on user interface software & technology*. 101–102.

[38] Bertil Gardell. 1979. *Psychosocial aspects of industrial production methods*. Department of psychology, University of Stockholm [Psykologiska inst ....

[39] A Ghoshal. 2017. Microsoft sued by employees who developed ptsd after reviewing disturbing content. *The next web* (2017).

[40] Etienne Grandjean and Wilhelm Hünting. 1977. Ergonomics of posture—review of various problems of standing and sitting posture. *Applied ergonomics* 8, 3 (1977), 135–140.

[41] Mary L Gray and Siddharth Suri. 2019. *Ghost work: how to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

[42] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. A data-driven analysis of workers' earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[43] Chien-Ju Ho and Jennifer Vaughan. 2012. Online task assignment in crowdsourcing markets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 26.

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

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

[46] Stephen E Humphrey, Jennifer D Nahrgang, and Frederick P Morgeson. 2007. Integrating motivational, social, and contextual work design features: a meta-analytic summary and theoretical extension of the work design literature. *Journal of applied psychology* 92, 5 (2007), 1332.

[47] Lilly C Irani and M Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 611–620.

[48] Lilly C Irani and M Six Silberman. 2016. Stories We Tell About Labor: Turkopticon and the Trouble with" Design". In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 4573–4586.

[49] V K. Chaithanya Manam, Dwarakanath Jampani, Mariam Zaim, Meng-Han Wu, and Alexander J. Quinn. 2019. TaskMate: A Mechanism to Improve the Quality of Instructions in Crowdsourcing. In *Companion Proceedings of The 2019 World Wide Web Conference*. 1121–1130.

[50] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

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

[52] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1301–1318.

[53] Ari Kobren, Chun How Tan, Panagiotis Ipeirotis, and Evgeniy Gabrilovich. 2015. Getting more for less: Optimized crowdsourcing with dynamic tasks and goals. In *Proceedings of the 24th international conference on world wide web*. 592–602.

[54] Michiel Kompier. 2002. The psychosocial work environment and health—what do we know and where should we go? *Scandinavian journal of work, environment & health* (2002), 1–4.

[55] Tage S Kristensen, Harald Hannerz, Annie Høgh, and Vilhelm Borg. 2005. The Copenhagen Psychosocial Questionnaire-a tool for the assessment and improvement of the psychosocial work environment. *Scandinavian journal of work, environment & health* (2005), 438–449.

[56] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25, 9 (2018), 1248–1258.

[57] Walter S Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F Allen, and Jeffrey P Bigham. 2013. Chorus: a crowd-powered conversational assistant. 151-162 pages. doi:10.1145/2501988.2502057

[58] Edith Law, Ming Yin, Joslin Goh, Kevin Chen, Michael A Terry, and Krzysztof Z Gajos. 2016. Curiosity killed the cat, but makes crowdwork better. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4098–4110.

[59] Richard J Loewenstein. 1991. An office mental status examination for complex chronic dissociative symptoms and multiple personality disorder. *Psychiatric Clinics* 14, 3 (1991), 567–604.

[60] Gale M Lucas, Albert Rizzo, Jonathan Gratch, Stefan Scherer, Giota Stratou, Jill Boberg, and Louis-Philippe Morency. 2017. Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI* 4 (2017), 51.

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

[62] Alwin Luttmann, Klaus-Helmut Schmidt, and Matthias Jäger. 2010. Working conditions, muscular activity and complaints of office workers. *International Journal of Industrial Ergonomics* 40, 5 (2010), 549–559.

[63] VK Manam and Alexander Quinn. 2018. Wingit: Efficient refinement of unclear task instructions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 6.

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

[65] Herbert Marcuse. 2013. *One-dimensional man: Studies in the ideology of advanced industrial society*. Routledge.

[66] Catherine C Marshall and Frank M Shipman. 2013. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *Proceedings of the 5th Annual ACM Web Science Conference*. 234–243.

[67] David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 224–235.

[68] Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. 2016. Taking a HIT: Designing around rejection, mistrust, risk, and workers' experiences in Amazon Mechanical Turk. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2271–2282.

[69] Adam S Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. 2016. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine* 176, 5 (2016), 619–625.

[70] Tanushree Mitra, Clayton J Hutto, and Eric Gilbert. 2015. Comparing person- and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1345–1354.

[71] Jesus Montero-Marin, Javier Garcia-Campayo, Mari Cruz Pérez-Yus, Edurne Zabaleta-del Olmo, and Pim Cuijpers. 2019. Meditation techniques v. relaxation therapies when treating anxiety: A meta-analytic review. *Psychological medicine* 49, 13 (2019), 2118–2133.

[72] K Murrell. 2012. *Ergonomics: Man in his working environment*. Springer Science & Business Media.

[73] Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*. 557–566.

[74] David Oleson, Alexander Sorokin, Greg Laughlin, Vaughn Hester, John Le, and Lukas Biewald. 2011. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*. Citeseer.

[75] World Health Organization. 2001. The World Health Report 2001: Mental health: new understanding, new hope. (2001).

[76] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153–163.

[77] Reinhard Pekrun, Thomas Goetz, Anne C Frenzel, Petra Barchfeld, and Raymond P Perry. 2011. Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary educational psychology* 36, 1 (2011), 36–48.

[78] Lisa M Perez, Jeremy Jones, David R Englert, and Daniel Sachau. 2010. Secondary traumatic stress and burnout among law enforcement investigators exposed to disturbing media images. *Journal of Police and Criminal Psychology* 25, 2 (2010), 113–124.

[79] Pierre Philip, Jean-Arthur Micoulaud-Franchi, Patricia Sagaspe, Etienne De Sevin, Jérôme Olive, Stéphanie Bioulac, and Alain Sauteraud. 2017. Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. *Scientific reports* 7, 1 (2017), 1–7.

[80] Judith J Prochaska, Erin A Vogel, Amy Chieng, Matthew Kendra, Michael Baiocchi, Sarah Pajarito, and Athena Robinson. 2021. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. *Journal of medical Internet research* 23, 3 (2021), e24850.

[81] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Estimating Conversational Styles in Conversational Microtask Crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 4 (2020), 1–23. Issue CSCW1.

[82] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Improving worker engagement through conversational microtask crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[83] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Just the Right Mood for HIT!. In *International Conference on Web Engineering*. Springer, 381–396.

[84] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Ticktalkturk: Conversational crowdsourcing made easy. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. 53–57.

[85] Priyanga Ranasinghe, Yashasvi S Perera, Dilusha A Lamabadusuriya, Supun Kulatunga, Naveen Jayawardana, Senaka Rajapakse, and Prasad Katulanda. 2011. 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 (2011), 1–9.

[86] Hyekyun Rhee, James Allen, Jennifer Mammen, and Mary Swift. 2014. Mobile phone-based asthma self-management aid for adolescents (mASMAA): a feasibility study. *Patient preference and adherence* 8 (2014), 63.

[87] Martin J Riedl, Gina M Masullo, and Kelsey N Whipple. 2020. The downsides of digital labor: Exploring the toll incivility takes on online comment moderators. *Computers in Human Behavior* 107 (2020), 106262.

[88] Senjuti Basu Roy, Ioanna Lykourentzou, Saravanan Thirumuruganathan, Sihem Amer-Yahia, and Gautam Das. 2015. Task assignment optimization in knowledge-intensive crowdsourcing. *The VLDB Journal* 24, 4 (2015), 467–491.

[89] Jeffrey Rzeszotarski and Aniket Kittur. 2012. CrowdScape: interactively visualizing user behavior and output. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 55–62.

[90] Jeffrey M Rzeszotarski and Aniket Kittur. 2011. Instrumenting the crowd: using implicit behavioral measures to predict task performance. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. 13–22.

[91] Susumu Saito, Chun-Wei Chiang, Saiph Savage, Teppei Nakano, Tetsunori Kobayashi, and Jeffrey P Bigham. 2019. TurkScanner: Predicting the hourly wage of microtasks. In *The World Wide Web Conference*. 3187–3193.

[92] Niloufar Salehi, Lilly C Irani, Michael S Bernstein, Ali Alkhatib, Eva Ogbe, and Kristy Milland. 2015. 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*. 1621–1630.

[93] Shruti Sannon and Dan Cosley. 2019. Privacy, Power, and Invisible Labor on Amazon Mechanical Turk. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[94] Saiph Savage, Chun Wei Chiang, Susumu Saito, Carlos Toxtli, and Jeffrey Bigham. 2020. Becoming the Super Turker: Increasing Wages via a Strategy from High Earning Workers. In *Proceedings of The Web Conference 2020*. 1241–1252.

[95] Saiph Savage, Andres Monroy-Hernandez, and Tobias Höllerer. 2016. Botivist: Calling volunteers to action using online bots. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 813–822.

[96] M Silberman and Lilly Irani. 2016. Operating an employer reputation system: lessons from Turkopticon, 2008-2015. *Comparative Labor Law & Policy Journal, Forthcoming* (2016).

[97] Rachel N Simons, Danna Gurari, and Kenneth R Fleischmann. 2020. "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 (2020), 1–26.

[98] Stacy L Smith and Edward Donnerstein. 1998. Harmful effects of exposure to media violence: Learning of aggression, emotional desensitization, and fear. In *Human aggression*. Elsevier, 167–202.

[99] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[100] Franklin Stein. 2001. Occupational stress, relaxation therapies, exercise and biofeedback. *Work* 17, 3 (2001), 235–245.

[101] Ryo Suzuki, Niloufar Salehi, Michelle S Lam, Juan C Marroquin, and Michael S Bernstein. 2016. Atelier: Repurposing expert crowdsourcing tasks as micro-internships. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2645–2656.

[102] Hiroki Tanaka, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. 2017. Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PloS one* 12, 8 (2017), e0182151.

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

[104] Syed H Tariq, Nina Tumosa, John T Chibnall, Mitchell H Perry III, and John E Morley. 2006. 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 (2006), 900–910.

[105] Jaime Teevan, Brent Hecht, Sonia Jaffe, and eds. 2021. *The New Future of Work: Research from Microsoft on the Impact of the Pandemic on Work Practices*. Technical Report. Microsoft. https://aka.ms/newfutureofwork

[106] Carlos Toxtli, Joel Chan, Walter S Lasecki, and Saiph Savage. 2018. Enabling Expert Critique with Chatbots and Micro Guidance. In *Collective Intelligence 2018*. 4.

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

[108] Philippe Verduyn, Iven Van Mechelen, and Francis Tuerlinckx. 2011. The relation between event processing and the duration of emotional experience. *Emotion* 11, 1 (2011), 20.

[109] Veniamin Veselovsky, Manoel Horta Ribeiro, Philip J Cozzolino, Andrew Gordon, David Rothschild, and Robert West. 2025. Prevalence and prevention of large language model use in crowd work. *Commun. ACM* 68, 3 (2025), 42–47.

[110] Andrew Vickers and Catherine Zollman. 1999. Hypnosis and relaxation therapies. *Bmj* 319, 7221 (1999), 1346–1349.

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

[112] JE Ware, Kristin K Snow, Mark Kosinski, and Barbara Gandek. 1993. SF-36 health survey. *Manual and interpretation guide. Boston: The Health Institute, New England Medical Center* (1993), 10–6.

[113] John E Ware Jr. 2000. SF-36 health survey update. *Spine* 25, 24 (2000), 3130–3139.

[114] Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.

[115] Mark E Whiting, Grant Hugh, and Michael S Bernstein. 2019. Fair work: Crowd work minimum wage with one line of code. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 197–206.

[116] Alex C Williams, Gloria Mark, Kristy Milland, Edward Lank, and Edith Law. 2019. The Perpetual Work Life of Crowdworkers: How Tooling Practices Increase Fragmentation in Crowdwork. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–28.

[117] EHC Woo, Peter White, and CWK Lai. 2016. Ergonomics standards and guidelines for computer workstation design and the impact on users' health–a review. *Ergonomics* 59, 3 (2016), 464–475.

[118] Meng-Han Wu and Alexander James Quinn. 2017. Confusing the crowd: Task instruction quality on amazon mechanical turk. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*.

[119] Luyan Xu, Xuan Zhou, and Ujwal Gadiraju. 2019. 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*. 1249–1252.

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

[121] Simone Zhang, Janet Xu, and A Alvero. 2024. Generative ai meets open-ended survey responses: Participant use of ai and homogenization.

[122] Ying Zhang, Xianghua Ding, and Ning Gu. 2018. Understanding fatigue and its impact in crowdsourcing. In *2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD))*. IEEE, 57–62.

[123] Mengdie Zhuang and Ujwal Gadiraju. 2019. In What Mood Are You Today? An Analysis of Crowd Workers' Mood, Performance and Engagement. In *Proceedings of the 10th ACM Conference on Web Science*. 373–382.