Understanding User Perceptions of Response Delays in Crowd-Powered Conversational Systems

TAHIR ABBAS, Eindhoven University of Technology, Netherlands
UJWAL GADIRAJU, Delft University of Technology, Netherlands
VASSILIS-JAVED KHAN, Digital Transformation, Sappi Europe, Belgium
PANOS MARKOPOULOS, Eindhoven University of Technology, Netherlands

Crowd-powered conversational systems (CPCS) are gaining considerable attention for their potential utility in a variety of application domains, for which automated conversational interfaces are still too limited. CPCS currently suffer from long response delays, which hampers their potential as conversational partners. The majority of prior work in this area has focused on demonstrating the feasibility of the approach and improving performance, while evaluation studies have primarily focused on response latency and ways to reduce it. Relatively little is currently known about how response delays in a CPCS can affect user experience. While the importance of reducing response latency is widely recognized in the broader field of human-computer interaction, little attention has been paid to how response quality, response delay, conversational context, and the complexity of the task affect how users experience the conversation, and how they perceive waiting for responses in particular. We conducted a between-subjects experiment ($N = 478$), to examine the influence of these four factors on the overall waiting experience of users. Results show that users 1) evaluated the waiting experience more negatively when the response delay was longer than 8 seconds, 2) underestimated the elapsed time but experienced more frustration in tasks with high complexity, 3) underestimated the elapsed time and experienced less frustration with high quality bot's utterances, 4) judged response delays to be slightly longer, and experienced more frustration in an emotion-centric CPCS compared to a task-centric CPCS. Our insights can inform the design of future CPCSs with regards to defining performance requirements and anticipating their potential impact on the user experience they can facilitate.

CCS Concepts: • **Human-centered computing** → **Natural language interfaces**; **Empirical studies in HCI**.

Additional Key Words and Phrases: Crowd-powered Conversational Agents, Response Latency, Perceived Response Latency, Time Perception, Waiting Tolerance, Cognition, Affect

ACM Reference Format:
Tahir Abbas, Ujwal Gadiraju, Vassilis-Javed Khan, and Panos Markopoulos. 2021. Understanding User Perceptions of Response Delays in Crowd-Powered Conversational Systems. In *CSCW '21: ACM conference on Computer Supported Cooperative and Social Computing, October 23–27, 2021, Virtual Conference.* ACM, New York, NY, USA, 42 pages.

1 INTRODUCTION

Conversational user interfaces have attracted considerable attention in human-computer interaction (HCI), already since the early days of this field, as they can be deployed quickly, they have low training costs and thereby can support a wide variety of applications, such as tutoring, companionship, and information retrieval, among others. Despite great progress, current artificial intelligence

CSCW '21, October 23–27, 2021, Virtual Conference

ACM ISBN XXX-1-XXXX-XXXX-X/XX/XX…$15.00

(AI) techniques and natural language processing are not yet capable of dealing with the full complexity of free-form conversational interactions, often resulting in conversation breakdowns [9]. Crowd-Powered Conversational Systems (CPCS) [48, 69] have been proposed as a remedy to these shortcomings of AI. CPCS can be more robust than current AI and can handle interaction with users in a fluid, multi-turn conversation. CPCSs make use of sophisticated recruiting, rewarding and user interface techniques to reduce latency of crowd input from hours to a few seconds [11]. A pioneering example of a CPCS is Chorus [69], which is a text-based conversational agent that assists end-users with information retrieval tasks by conversing with an online group of workers in real time. Since such CPCSs rely (fully [69] or partially [48]) on operation by humans, significant response delays can be expected that might frustrate end-users and compromise the overall user experience. The typical workflow of CPCS is depicted in Figure 1, highlighting locations where latency may occur.

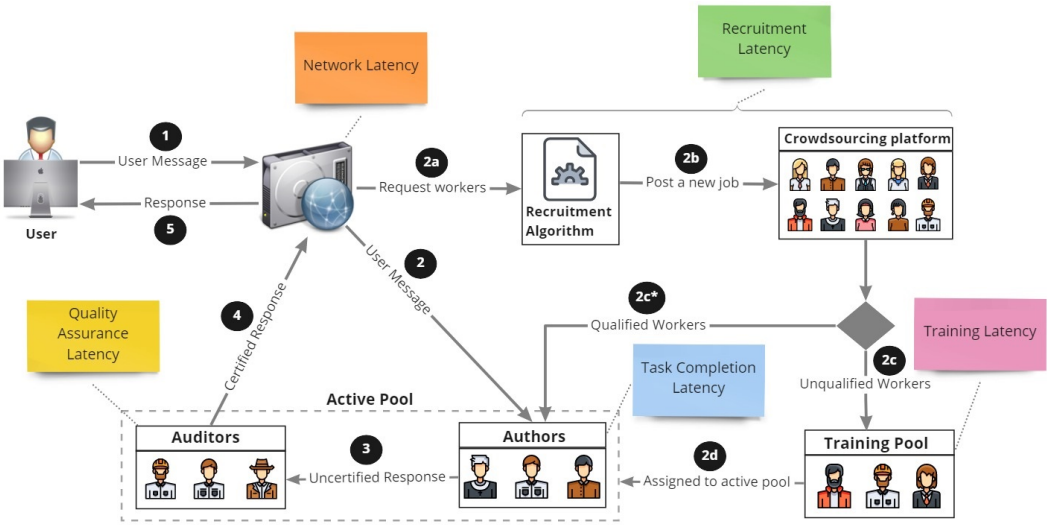

Fig. 1. This diagram depicts a typical workflow of CPCS. (1) First, user sends a message to the server. This message may take some time to reach the crowd workers depending on network bandwidth, resulting in *network latency* [67]; (2) The message is then sent to a pool of active workers; (2a-2b) if there are no workers in the active pool when user sends a message, then server requests recruitment algorithm to hire a sufficient number of workers from the crowdsourcing platform. The time it takes for an interested crowd worker to accept a newly posted task is referred to as *recruitment latency* [43]; (2c) When workers take their first task, they are often required to complete tutorials or qualification tasks before being able to perform actual tasks. This causes *training latency* [43]; (2d) After training, they are assigned to a pool of active workers; (2c*) If they are already qualified for the job, they can begin immediately, albeit this is rare. Once in the active pool, they are allocated to one of two roles: one set of workers (authors) creates responses, while another group of workers (auditors) validates those responses before they are returned to the users. The *task completion latency* is the time required for a worker to finish a task, which varies according to the workers' skill, fatigue, and a variety of other factors [52]. (3) The messages generated by one group of workers (authors) are forwarded to another group of workers (auditors) for validation, introducing *quality assurance latency*; (4) This certified message is then transmitted to the server, where it may encounter delays owing to network latency; (5) finally, the message is reached to the intended user.

Research on time perception in HCI suggests that with longer response time, users show more frustration and stress [104], perform poorly [18], demonstrate dissatisfaction with the system [27]

and may even be affected physiologically [98] (e.g., elevated blood pressure). Although there is wide agreement concerning the negative effects of long response times on end-users, views differ on what waiting times are acceptable; some researchers advocate 10 seconds as an upper limit [85], while others suggest a stricter 8 seconds limit [35]. The notion of an acceptable time delay refers directly to how users experience delays rather than on the objectively measurable time delay between user utterances and system responses. The majority of work in CPCS has only focused on reporting or reducing the "actual" or system response latency (See section 2.2 for more details), but relatively little is known concerning the subjective experience of such response delays. Here we examine the "perceived" latency, which is a subjective measure pertaining to the user experience which pertains to the perception of the waiting time between conversational turns. Evaluating the perceived waiting times in CPCS and how they influence user's attitudes towards the system can help prioritize requirements and guide the design of these systems, leading to our first research question:

> **RQ1**: *How do end users perceive the response latency of a CPCS while interacting with it?*

One can expect to receive responses of mixed quality from CPCSs. This is in part due to incongruity of workers' skills and expertise with the task, biased interests, unfair incentives, poorly designed tasks, insufficient workers' training, among others [3, 6, 26, 61]. A high-quality response is one that is appropriate to the user's utterance, contains helpful and concrete advice and gives individualized attention to the user [123]. We hypothesize that a high-quality response from CPCS can positively influence the overall waiting experience. For this paper, *waiting experience* refers to perceived waiting time (believing that less or more time has passed than the actual duration), cognitive appraisals of the wait [70, 93, 116] (i.e., evaluations based on instrumental beliefs, such as performance and usefulness), affective appraisals of the wait [70, 93, 116] (i.e., evaluations based on feelings, emotions, and moods) and perception of overall interaction quality. We refer to the cognitive and affective appraisals collectively as *appraisals of the wait*. Our selection of measures for the appraisals of the wait is based on prior research in time perceptions that has shown an inverse relationship between utilitarian evaluations (e.g. performance and lengthy waiting time [18]) and emotional states (e.g. stress and lengthy waiting time) [104], and between the quality of the response contents and negative affect states [125] . Thus, in the case of high-quality responses, users may perceive time to pass faster, may experience less stress, and a better overall interaction quality. This line of reasoning leads us to the following research question:

> **RQ2**: *How does the quality of responses generated by a CPCS influence the waiting experience of users?*

Recently, researchers have adopted machine learning to quantify task clarity [37] and task complexity [124] for crowdsourcing tasks and studied their impact on the worker performance [36]. In our study, we want to examine whether the complexity of a conversational task has any impact on the time perceptions within the CPCS context. For instance, take an example of an information retrieval based CPCS where a simple search task requires only retrieving factual information from an information resource (e.g., What is the name of the deepest point in the ocean?). Next to that, a more complex search task involves also organizing factual information into a new pattern or generating a new plan (e.g., Help me to put together two thirty-minute low impact exercise programs for my 90-year-old granny). Similarly, a stress management CPCS requires answering user complaints involving complicated and multi-layered stressors and entails higher cognitive complexity and time investment than a simple stressor concerning weight loss or exam stress. Correspondingly, workers may need more time in generating a response for a complex case and the user might expect the CPCS to take longer to respond. Arguably, this could

influence the overall waiting experience positively, with users underestimating the elapsed time and experience comparatively less affective responses to the delay when the complexity of task is high. This notion is supported by prior research that explains that increase in cognitive load can decrease the perceived time delay [5, 13]. It should be emphasized that task complexity can be related to both the difficulty of the task [122] and the context of the conversation [107] (see next paragraph). Since one of the study's overarching aims was to gain a better understanding of the influence of task complexity as a broad construct, we employed two types of tasks to reflect complexity in terms of the conversational context in addition to the complexity specified at the task level, as opposed to considering either of these alone. In the context of CPCSs this leads to the following research question:

> **RQ3**: *How do conversational tasks of varying complexity affect the waiting experience of users in CPCS?*

Furthermore, we wanted to explore the effects of two different task types on the waiting experience for the following three reasons: (1) We created two distinct task types in light of a prior research indicating that user expectations regarding delays may vary according to the type of conversational context [89, 107]. For instance, in our recent study [1], we evaluated the usefulness of several engaging interventions or waiting-time fillers in lessening the impression of time. We discovered that fillers, notably those that demand users to complete a secondary activity while waiting, were more helpful in Information Retrieval (IR) activities than the stress mitigation task. Likewise, when prolonged delays in the service of a robot are inevitable, the research [89] found that consumers in the chitchat settings were marginally less satisfied with lengthier waits than users in the IR conditions. Thus, we were intrigued if this distinction remained true when we enabled end users to interact with various sorts of activities in CPCS, when no fillers were offered. This knowledge would aid the developer of CPCS in determining the upper and lower limits of an acceptable waiting time in CPCS for different task types and in creating applicable mitigation measures depending on the conversational context. (2) Furthermore, we focus on both stress mitigation and IR tasks due to the pervasiveness of these tasks in contemporary CPCSs. For instance, Panoply [83] and CoZ [3] are crowd-powered conversational systems that provide on-demand emotional support to people. In the context of IR, Chorus [69] and Evorus [48] are crowd-powered text-based chatbots that assist users with IR tasks. The two tasks differ in breadth and depth of focus, duration, and speed [42]. In stress mitigation, workers engage with users through multi-turn conversation by asking focusing or evoking questions, expressing empathy, and reflecting and validating the users' concerns. Therefore, users are expected to experience shorter delays, giving the impression of a continuous and fluent dialogue, enhancing therapeutic engagement, and sustained usage. On the other hand, IR tasks can be completed with either single or fewer dialogue exchanges with considerable delays because workers sometimes have to find the relevant information on the Web. In Table 1, we have provided some examples of task-centric and emotion-centric dialogues to help readers distinguish between the two tasks. (3) Finally, we choose to explicitly add the stress mitigation task because of its ubiquitous relevance. While technology-assisted behavioral interventions can be implemented more easily than in-person interventions, they are limited in their ability to comprehend or express emotions and provide contextualized assistance [62, 87]. As a result, researchers have lately begun to examine crowd-sourced emotional support solutions [83]. However, the bulk of existing tools are oriented on the asynchronous approach, which places a low premium on response time. However, research indicates that when creating conversational bots for mental health applications, response time is a crucial characteristic to consider [23]. Thus, we sought to ascertain the acceptable waiting time for real-time crowd-powered affective applications,

which are gaining traction in response to global mental health challenges. This potential influence of task type motivates the following research question:

RQ4: *How does the nature of the conversational task affect the waiting experience of user in CPCS?*

Table 1. Examples of task-centered and emotion-centered dialogues

	IR based bots	Examples		Topics of Motivational Interviewing (MI) / Active listening	Examples [88, 120]
Task-centric Dialogue	Chorus [69]	Hi, I want to know how to make lasagna?	Emotion-centric Dialogue	Focusing Question	How would you feel about that?
	Evorus [48]	Can you find me some good restaurants in Pittsburgh?		Evoking Question	What worries you about your current situation?
	Guardian [47]	hello, I like to know some information about the movie Titanic		Reflective listening	If I heard you correctly, this is what I think you are saying: *.
	Humorous Bot [19]	I want to understand how rocks are formed		Expressing Empathy / MI-adherent statement	Anyone would feel this way if they were in your situation.

To answer **RQ4**, we developed two chatbots (Figure 2 & 3) to simulate the CPCS for two different contexts: stress mitigation and IR. For stress mitigation, we developed *Stress-bot*, which is a chatbot we built on the principle of active listening [100] and Motivational Interviewing (MI) skills [78]. To support IR tasks, we built an *IR-bot*, which is a chatbot that answers IR queries in the health, science, and entertainment domains. To answer **RQ1**, we varied the response time of the two chatbots based on the following geometric sequence: 2s, 4s, 8s, and 16s. We adopted this sequence from prior research in computer response theory and crowdsourcing [3, 18]. To answer **RQ3** we developed two variants of the bots each supporting a different complexity of conversational tasks: a high and a low complexity. For Stress-bot, the high complexity task involves a challenging scenario with multiple input stressors (any life event or past experience that causes stress) whereas the low complexity task involves a single stressor and a relatively simple situation, such as exam stress. For an IR-bot, the complexity of the search task was varied based on the cognitive complexity framework proposed by Kelly et al. [57] which provides different reusable sample search tasks that we used in the current study. To address **RQ2**, we made two versions of bots based on response quality: high and low. For the Stress-bot, a high-quality response was based on statements that were vetted by professional therapists whilst a low-quality response was simply based on small-talk. For an IR-bot, a high-quality response was detailed, complete and meaningful, while a low-quality response was incomplete and contained an element of hesitancy or uncertainty. In total we developed 32 experimental conditions: **2 x Context, 2 x Complexity, 2 x Quality and 4 x Delay**.

We conducted a between-subjects ($N = 478$) experiment on the Prolific crowdsourcing platform to study the influence of four factors (described above) on the waiting experience. The waiting experience was assessed by the perceived waiting time (in seconds), an appraisal of the wait (long/short judgments and affective responses), and the interaction quality. Our findings show that:

- **RQ1**: Among the four levels of delay compared in the experiment, the 16s delay resulted in more irritation and boredom in the participants. The 8s delay was considered tolerable while the 2-4s delay was perceived as robotic and unrealistically fast.
- **RQ2**: When the quality of the response content was high, participants reported a more positive waiting experience and showed a greater engagement with the CPCS.
- **RQ3**: When the complexity of the task was low, participants overestimated the elapsed time than the actual duration, and perceived quality of the response more positively.

- **RQ4**: When interacting in information retrieval tasks (IR-bot), participants underestimated the elapsed time, experienced less stress, and perceived the quality of interaction more positively than participants who performed the stress mitigation tasks (Stress-bot).

The remainder of this article is structured as follows. In Section 2, we review related work. In Section 3, we detail our methodology including the design of the conversational interfaces, the procedure and the measures used in the study. In Section 4, we detail our results including participants' feedback on long delays. In Section 5, we discuss the insights gained from our results. Finally, in Section 6, we conclude with some implications for designing future CPCSs.

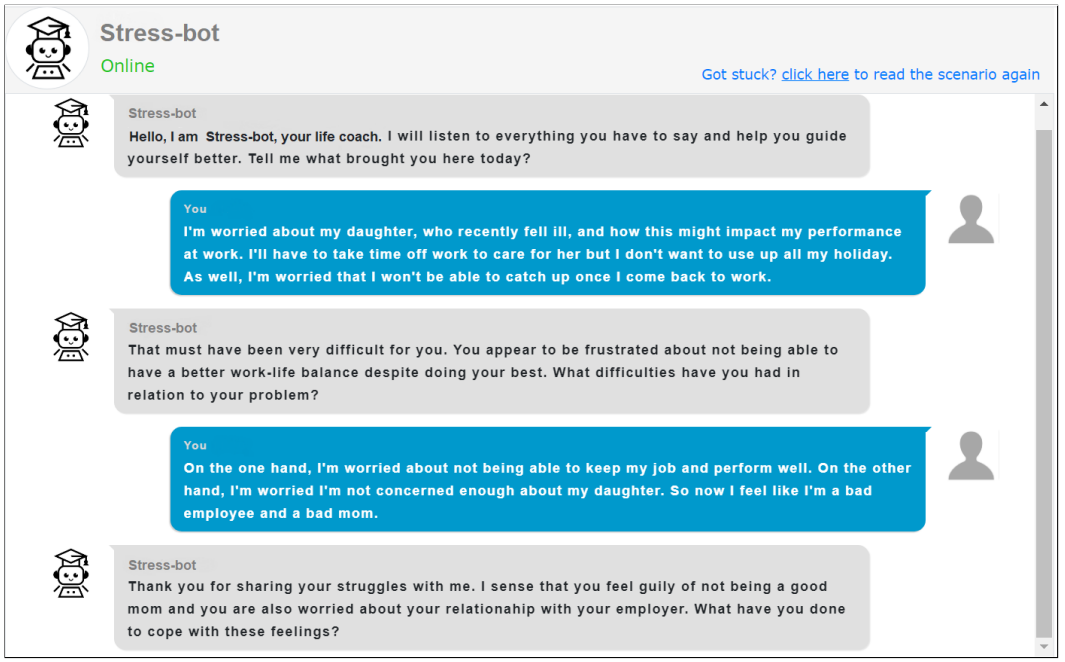

Fig. 2. User Interface of Stress-bot. This figure shows a portion of real chat between Stress-bot and one of the participants from an experimental condition where task complexity and quality of responses of Stress-bot were set to high. Stress-bot initiates the discussion and then allows users to share their ostensible worries based on the description of fictional character. Participants can access the description of the character anytime by clicking the "Got stuck?" link; after that, the description of the character is shown in a popup window.

2 RELATED WORK

2.1 Psychology of System Response Time in Human-Computer Interaction

Empirical evidence appears to confirm the notion that system response time (SRT) is an important factor that influences many aspects of the interaction, such as user satisfaction [27], performance [18], user stress [104], quality of work [53], among others. For instance, Davis and Heineke [27] found evidence that high values of perceived waiting time and actual waiting time can reduce the customer satisfaction in a service operation context. A study of response delays in browser-based applications [44] found that for every three second increase in the system response time, there is an average of .22 drop in average satisfaction and suggested that dissatisfaction may lead to discontinuing the use of the application. Toomim et al. [114] found that Mechanical Turk workers

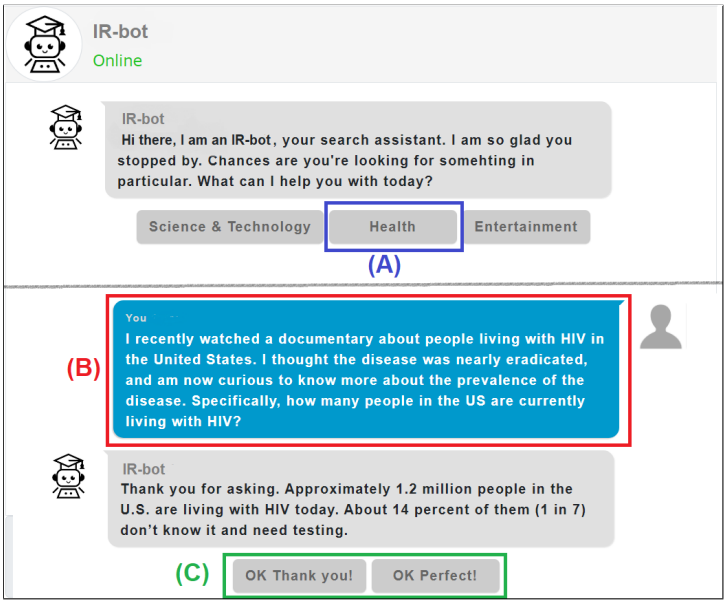

Fig. 3. User Interface of an IR-bot. (A): An IR-bot initiates the conversation by displaying some predefined domains in the form of quick replies. These quick replies disappear when the user chooses one of them.(B): After selecting a domain of interest, an IR-bot then replaces quick replies with a predefined search task, and displays it in a chat bubble. (C): Finally, users can click a quick reply to exit out of the IR task.

preferred to continue with the task when the time to complete the task was lowered as predicted by Fitts' Law. Jacko et al. [53] investigated the combined influence of network delay (short, medium, or long) and media (text or text/graphics) on usability assessment. One important finding of their experiment was that participants perceived the quality of the document to be lower when the delays were lengthier even when the text was augmented with graphics. Some researchers have studied the effect of long delays on the affective experience concerning SRT, such as frustration and stress. For example, Schleifer and Amick III [104] claimed that participants who experienced longer SRTs provided higher ratings of frustration, impatience, and stress. Other researchers studied the physiological effects (rapid heart rate and blood pressure) associated with longer SRTs. Seen in this light, Riedl and Fischer [98] found that the heart rate and blood pressure of those participants who experienced longer SRTs was elevated.

2.1.1 Acceptable System Response Time. Earlier research results pertaining to what response delays are acceptable for users diverge [75]. Shneiderman [109] stated that a system that responds in under two seconds is considered more acceptable than a system that takes longer to respond. Nielsen in his book of Usability Engineering [85] argued that "10 seconds is about the limit for keeping the user's attention focused on the dialogue." Fröhlich [35] provided empirical evidence that while interacting with a speech application, users can only tolerate silences not exceeding 8s. Another study [54] examined the effects of delayed system responses upon user stress and the mental strain they feel. Results show that 63% of their participants wanted a response delay of maximum 5 seconds, while only 20% were positively inclined towards 10-seconds delay.

2.1.2 Response Time of Chatbots. In human-chatbot interaction, only a few studies investigated the effects of variable response delays on the perceived interaction quality. Such studies mainly relate

the response times of chatbots with their perceived humanness and persuasiveness. Stemming from the work of persuasion in computer mediated communication, Moon [80] found a non-monotonic relationship between response latency and the persuasiveness of a chatbot's messages. Participants were invited to solve the desert survival problem together with a chatbot. Participants who interacted with a chatbot with moderate response latency (5-10s) were more willing to modify their preferences according the chatbot's suggestions, compared to those who interacted with a chatbot that responded with either long delays (13-18s) or short delays (0-1s). Gnewuch et al. [39] studied the relationship between response delays of a chatbot and its perceived humanness and social presence. The authors dynamically varied the response delays based on the complexity of the message (calculated with the Flesch-Kincaid grade level, a readability test to determine how difficult a message is to understand). Their results indicate that a chatbot that sent dynamically delayed responses was perceived to be more human-like and had more social presence than a chatbot that sent messages quickly without significant delays.

2.2 Response Delays in Crowd-powered Conversational System

2.2.1 Real-time Crowdsourcing and its Applications. In its early days, crowdsourcing was treated as an asynchronous mechanism by which tasks could be completed in batches, taking hours to complete. In recent years, researchers have developed methods to recruit workers on demand from crowdsourcing platforms and created techniques to reduce latency. Such methods enable real-time crowdsourcing (RTC) [66] and the resulting systems that support the RTC concept are known as real-time crowd-powered systems. Real-time crowd-powered systems use recruiting, rewarding and user interface techniques to reduce the latency of crowd input from hours to a few seconds [11]. VizWiz [12] is one of the earliest real-time crowd-powered system which helps blind users in answering visual questions about their surroundings by sending photos of objects and audio questions from their phones to crowd workers. To make the workers available on demand, VizWiz incorporates the quikTurkit algorithm that hires workers in advance and keeps them engaged with some prior tasks. Adrenaline [11] is a mobile application that leverages crowd intelligence to quickly select the best frame in a video, taking less than 2 seconds to do so in a ten-second video. Adrenaline implements retainer model where workers are paid a fixed fee for waiting and a small reward (3 cents) for quickly responding to alerts. Legion [67] allows end-users to crowdsource the real-time synchronous control of the locomotion of a robot through keyboard commands, while observing a real-time video feed. Legion:Scribe [65] is a crowd powered system that leverages the services of unskilled crowd workers to provide captions to deaf people in real time.

2.2.2 Response Delays in CPCS. For the reader's convenience, we have summarized the reported latencies of current CPCSs in Table 2. A pioneering example of a crowd powered conversational agent is Chorus [69], which supports information retrieval. Chorus lets workers propose responses and vote on the responses of other workers in order to support a consistent dialogue. It implements a working memory model to sustain a coherent conversation over time. Chorus was found to have a response latency of 103.4s with a single worker, which however could be reduced down to 44.6s by employing multiple workers and choosing among their inputs with a voting mechanism. Since the main focus of that work was maintaining a consistent and sustained dialogue with multiple workers, authors did not examine the perceived latency of users. Chorus:view [68] is a conversational application, which was built on top of Chorus. It assists visually impaired users to ask general questions about their surroundings by engaging them in a continuous conversation with crowd workers. Chorus:View broadcasts real-time video stream and audio from the users' mobile phones to crowd workers allowing them to access the conversational feature implemented with Chorus. In a trial with blind users, Chorus:View was able to accomplish various tasks with

Table 2. Summary of Response Latencies in Crowd-powered Conversational systems. None of the studies have looked into the subjective waiting experience in CPCS except reporting system response latencies.

System	Average Response latency
Chorus [69]	Single Worker: 103.4s Suggest Condition: 44.6s Filter Condition: 50.13s
Chorus:View [68]	Product Detail Task: 295s Information Finding Task: 351.2s Navigational Task: 182.3s
RegionSpeak [127]	38s to 2 minutes
Guardian* [47]	Yelp Search: ~70.04s Rotten Tomatoes: ~48.6s Weather Underground: ~140.1s
Entity Extraction System [50]	40.95s
InstructableCrowd+ [49]	Crowd-only: ~5 minutes Crowd voting: 20 minutes
Crowd of Oz [2]	Single Worker: 8.82s Group of two or four: 6.79s Group of eight: 4.12s
CRQA [103]	50s
Evorus [48]	Not Reported

*Total response time was inferred by adding parameter filled time and JSON filled time based on table 3 from [47]

+Total response time was inferred based on the average time users spent in conversing with the crowd and total time it took for a crowd to create a rule

an average response time of 295s, 351.2s and 182.3s for product detail, information finding and navigational tasks, respectively. However, they also did not examine how the blind users perceived the response delays. RegionSpeak [127] is an advanced tool built on top of VizWiz [12] and Chorus [69], which combines Panoramic images (by combining multiple photographic images) and parallel labeling approaches with multiple workers to label multiple objects in a scene. As a result, visually impaired users can conveniently explore the spatial layout of the objects in the space. A user test found that the average response time of RegionSpeak ranged from 38s to ~2 minutes.

Guardian [47] is a crowd powered spoken dialogue system that combines Web APIs with crowd-sourcing to enhance the scope of open dialogue systems in two phases. In the first offline phase, question and answer (QA) pairs are collected with non-expert workers for each API and then match each QA pair with the related parameter. Subsequently, in the online phase, non-expert workers extract parameter values from the running dialogue to obtain the JSON response from the web API service. After receiving the JSON response they translate it into a natural language answer understandable by the user. The performance of Guardian was described in terms of the time needed to fill a given parameter (parameter filled time) and the time needed to acquire the JSON string from the web API (JSON filled time) since the end of the user's first utterance input. Thus, the total time for receiving the answers to the users' questions is the sum of two values. For instance, the time it took to receive the correct response was found to be 70.04s for Yelp search API, 48.6s for Rotten Tomatoes and 140.1s for Weather Underground (cf. Table 3 from [47]).

Huang et al. [50] developed a crowd-powered system to extract entities from a running dialogue in real time using an ESP game. Unlike an ESP game where players are shown images to label, pairs of workers were shown a complete dialog and a description of an entity that they needed

to extract from the running dialog. When both workers would identify the same entity, then the task would be considered complete. An empirical evaluation showed that aggregating answers from workers yielded good quality results in less than a few seconds. The first worker arrival time since the publishing of MTurk task was 30.83s, the first response took 37.14s to arrive, while the first matched response took 40.95s on average. They did not examine the workers' self-reported perceived latency.

InstructableCrowd [49] is a conversational system which allows end users to communicate their requirements regarding IF-THEN rules to a group of workers. In return, crowd workers assist end-users by creating IF-THEN rules that then run on their mobile phones. Users can also review, edit, and approve these rules. In their experiment, user took on average 3 minutes and 45 seconds to converse with the crowd concerning the rules and system took another 1 minute to create a rule as per users' demands, resulting in ~5 minutes delay in total. In a condition where they merged rules from multiple workers, it took approximately 20 minutes for a system to receive rules from all workers.

Crowd of Oz (CoZ) [2] is a crowd powered conversational system for social robotics that can outsource conversational tasks of the Pepper robot to a synchronous group of workers. CoZ transcribes the speech of the user and forwards it to crowd workers along with audio-video (AV) feed captured through the camera and microphones of Pepper. Using a web interface that displays AV feeds and a chat box, crowd workers compose a text message, which is then spoken out by the Pepper robot. In the evaluation, authors systematically varied the number of workers who simultaneously handle the speech of the robot for a stress mitigation task, and studied its effects on the response latency and response quality. Their experiment showed response delays of 8.82s, 6.79s, 6.79s and 4.12s with one, two, four and eight workers, respectively without effecting the quality of responses adversely. However, they did not examine the perceived latency for users.

Some researchers have combined crowdsourcing and computational intelligence to automate CPCSs. For instance, CRQA [103] is a hybrid crowd powered question answering system for informational tasks. It includes a crowdsourcing module to validate the answers generated by CRQA from existing data sources or optionally to ask workers to provide an answer if CRQA fails to generate relevant answer. Once the question is posted to CRQA, it allows workers 50s to respond. After the crowd input is received, CRQA employs a learning-to-rank model to select the final answer. Similar to CRQA, Evorus [48] also follows a hybrid approach to generate and validate candidate answers from crowd workers. Unlike CRQA that relies on existing information sources, it incorporates existing chatbots though REST APIs. Over time, it learns to select the high-quality answer from chatbots based on workers' past evaluations of answers generated by chatbots.

In summary, while a considerable body of research in CPCS has examined the actual latency and ways to reduce it, less attention has been paid to the latency as it is perceived by users of CPCS and the underlying factors that contribute to their waiting experience. Our study is aimed at filling this important gap in the literature.

3 METHOD

We conducted a between-subjects experiment on the Prolific crowdsourcing platform to study the influence of four factors of CPCS (Context, Complexity, Quality and Delay) upon the waiting experience of users (see section 3.8 for more details about the dependent variables). Specifically, we compared 32 different conditions: 2 (*context*) × 2 (*complexity*) × 2 (*quality*) x 4 (*delay*). We designed a chatbot for each context (stress and IR). The user interface of these bots was designed using the TickTalkTurk[1] library by Qiu et al. [96]. The server application was developed using Flask, a

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

Python-based web framework[2]. The ethics review board at the [anonymous university] approved the methods presented in this paper.

3.1 Rationale For Choosing Different Levels of Delays

Based on a recent study [2], which examined how varying the number of workers (N=1,2,4,8) engaged concurrently with a task influences the response latency and the quality of the responses, we considered the delays of 4 and 8 seconds as representative of the performance range for real time crowdsourcing [2]. To account for slower crowd work and to introduce sufficient variation we included a delay of 16s as the longest duration considered in this experiment.

Prior research in HCI indicates that the system should respond within two seconds after receiving input [108]. This rule is frequently used in HCI research as a design guideline. We wanted to determine if this rule still applied to hybrid intelligence driven CPCS, where responses are expected to arrive faster than when only humans are involved. Furthermore, this 2s condition functioned as a baseline against which contemporary hybrid CPCSs like Evorus [48] could be compared. Furthermore, participants were made aware that, despite the response being received more quickly in 2s condition, human corroborators were still involved to evaluate the responses, as we have specified in the instructions (Cf. Section 3.7)

By this reasoning we arrive at the following geometric sequence of the delays: 2s, 4s, 8s, and 16s. A similar geometric sequence was also used in a classic study on user interface response delays [18] which examined the relationship between computer response times (2, 4, 8, 16 and 32s) and user performance with simple data entry tasks.

3.2 Design of the Stress Management Conversational Task

We focus on stress management task since stress-related disorders are becoming more prevalent and are connected with significant health concerns [25]. While psychotherapies have been shown to reduce stress, the vast majority of the world's population continues to lack access to psychologists [119]. To reach the widest possible audience, we need accessible, inexpensive, and anonymous mental health consultations. Due to the inability of artificial intelligence (AI) to give answers for users' particular circumstances, AI-based mental health interventions currently have a high rate of attrition and low adherence [72]. As a result, it is critical to develop a crowdsourced psychological intervention capable of resolving these challenges. Given its importance, some projects [55, 83] integrating affective computing and crowdsourcing – dubbed affective crowdsourcing [82] – began to emerge more recently. However, prolonged delays may obstruct their applications. Additionally, we added the stress task because of its open-ended nature. In general, the dialogues of such systems are either open-ended (there is no specific answer, and the dialogue is intrinsically useful) or closed-ended (i.e., the dialogue has a specific answer). We used the IR task to address the closed-ended dialogue case. Additionally, we wanted to add an open-ended case, which is why we included the stress-related task. Thus, examining the delays in such dialogues would offer designers with a plethora of information about the maximum permissible delay, mitigation measures for reducing perceived latency, and the most efficient use of available human resources for stress-related conversations in CPCS.

To design the stress management conversational tasks, we first collected videos clips from YouTube (based on the strategy described by Pérez-Rosas et al. [90]) , which demonstrate simulated counselling sessions between a client and a counselor using motivational interviewing [115]. We examined videos that demonstrate different situations, such as controlling anger, exam stress, social anxiety, relationship problems, among others. Finally, we selected two videos that fulfill our

[2]https://flask.palletsprojects.com/en/1.1.x/

criterion of complexity (see section 3.2.1 for more details). We then transcribed these videos with the assistance of Otter.ai platform[3], which is an advanced AI-powered transcription platform that generates very accurate machine-generated text from voice conversations. The generated text is further augmented with speakers' tags. This feature of Otter helped us to differentiate between the client and counselor's utterances.

In the simulated CPCS, workers were supposed to choose the predefined or template-based user utterances rather than typing their own responses. We preferred this set up over allowing users to converse "freely" and "openly" with the simulated CPCS in order to allow a more controlled experiment. The alternatives of conversing freely with an error prone artificial agent or an actual CPCS would introduce several confounding factors regarding the direction and content of the conversation and the variability in response rates. Thus, we asked participants to role play predefined characters, and carry out dialogues with the simulated stress-based CPCS. An example of a fictional character from a simple stress task is presented in Appendix A. Full transcripts of a few dialogues from different experimental conditions and other project related details can be found at this anonymous URL: https://bit.ly/3uPijLb

The simulated CPCS was designed in the form of a chatbot that was customized to respond based on the description of the character in the narrative. One way for a conversational agent to help users resolve their negative feelings is to invite them to share thoughts that stress them with the agent, by assuming the role of an active listener. In this active listening mode, the agent lets users open up and talk about her problems without giving any specific resolution or guidance for dealing with their problems. The agent has a set of predefined utterances or questions it can present to the user. This strategy is also used in Woebot[4], which is a very popular mental health chatbot. To simulate stress management tasks with CPCS, we developed Stress-bot. Stress-bot is simply a chatbot that we built on the Rogerian principle of active listening [100]. We will further expand on how Stress-bot works in the section 3.7. In the next two sections, we explain on how we handle the complexity and the response quality of Stress-bot.

3.2.1 Complexity of the Stress Task. To investigate the influence of complexity for the stress task, we created two variants of CPCS based on Wood's pioneering work [122]. Wood [122] suggested a task model with three critical components: products, acts, and information cues. We created two levels of complexity based on the dimension of static complexity, which is related to task design. It refers to the number of distinct acts and cues required to complete the task. To create tasks with a static complexity dimension, we make our decision based on the number of stressors present in the stress task. For example, Ruscio and Holohan [102] provided a number of characteristics that can be used to characterize the complexity of a therapeutic stress case. They characterized one of the criteria as multiple, significant current stressors. Thus, the presence of additional stressors indicates that the case is more difficult to resolve. Each stressor can be mapped with a static complexity component that necessitates the execution of specific acts and information cues. For example, in a complex case, we chose a video[5] in which the client stated two stressors: concern for her daughter, who suddenly became unwell, and concern about her performance at the company where she recently began working. Both types of stressors demand specific acts (e.g., describing stressors, feelings, causes, coping strategies) and information cues (recalling number of facts from the description of a fictional character to support a certain act) to execute the task. Thus, more stressors indicate that a task would place greater cognitive demands on the individual, thereby

[3]https://otter.ai/
[4]https://woebothealth.com/
[5]https://youtu.be/osROod3Hmpg

increasing its complexity. For a simple case, we chose a video[6] in which a client has a single source of stress: concerns about school performance.

3.3 Design of the Information Retrieval Conversational Task

In order to simulate IR tasks with CPCS, we built an IR-bot. An IR-bot is a chatbot designed to answer IR queries. We did not allow users to type open queries while interacting with an IR-bot. Instead, we provided custom keyboard options or predefined reply options following the guidelines of cognitive complexity framework [57]. We explain further how an IR-bot works in the section 3.7.

For the design of information retrieval tasks, we relied on the framework developed by Kelly et al. [57] based on Bloom's taxonomy which distinguishes between six types of cognitive processes: *remember, understand, apply, analyze, evaluate, and create.* These cognitive processes are ordered based on the amount of effort required to execute them. For instance, *remembering* tasks are the simplest in information retrieval, requiring the user to just identify or recognize a fact from an information source (e.g., "how many people in the US are currently living with HIV?"). On the other hand, *create* tasks are the most complex tasks, requiring the user to compose a complete plan by putting elements together or reorganizing elements to create something new (e.g., "Identify some basic designs that you might use and create a basic plan for constructing the derby car"). So, the more complex information retrieval task was based on the *create* principle while the simple task was created based on the *remember* principle. Kelly et al. [57] has already developed a dataset of reusable search tasks based on the cognitive complexity framework. We chose three different *remember* tasks, one for each of the three different domains: *Health, Science & Technology and Entertainment* and three *create* tasks for the same domains. Hence, our final corpus contains three simple and three complex IR tasks, as shown in Table 3.

Table 3. Examples of Search Tasks, adopted from Kelly et al. [57]

Domain	Simple Case	Complex Case
Health	I recently watched a documentary about people living with HIV in the United States. I thought the disease was nearly eradicated, and am now curious to know more about the prevalence of the disease. Specifically, how many people in the US are currently living with HIV?	My great granny's doctor has told her that getting more exercise will increase her fitness and help her avoid injuries. I need an exercise program for her. She is 90-years old. Help me to put together two thirty-minute low impact exercise programs that she could alternate between during the week.
Science & Technology	I recently watched a show on the Discovery Channel, about fish that can live so deep in the ocean that they're in darkness most or all of the time. This made me more curious about the deepest point in the ocean. What is the name of the deepest point in the ocean?	After the NASCAR season opened this year, my niece became really interested in soapbox derby racing. Since her parents are both really busy, I agreed to help her build a car so that she can enter a local race. The first step is to figure out how to build a car. Identify some basic designs that you might use and help me to create a basic plan for constructing the car.
Entertainment	I recently attended an outdoor music festival and heard a band called Wolf Parade. I really enjoyed the band and want to purchase their latest album. What is the name of their latest (full-length) album?	My local Triple-A affiliate baseball team has decided that it is time for a new mascot and are holding a contest where fans can enter suggestions. Being a loyal fan, I have decided to enter the contest. I want to suggest a mascot that will appeal to many people and will represent important qualities of a baseball team. The team is a part of the International League, so I want to avoid suggesting a mascot that is already represented in this league. Which mascot would I pick and why?

[6] https://youtu.be/KRDH89uP8wI

3.4 Operationalizing the Quality of Responses for Stress-bot and IR-bot

In the current study, a high-quality responses of Stress-bot and IR-bot met three criteria provided by previous research in the field of chatbot interfaces [123]: Appropriateness, empathy, and helpfulness. Appropriateness refers to the response that is more relevant to the user request; empathy means the response should provide each user personalized attention and make them feel valued; Finally, the term "helpful" refers to a response that provides relevant and specific advice in response to a user's request and free of any ambiguity. These criteria were not met by a poor-quality response. We made sure that a high-quality response is substantially different than a low-quality response.

Furthermore, high quality responses of Stress-bot were constructed using motivation interviewing principles, which were vetted by psychologists. For designing the high quality templated responses of Stress-bot, we adopted a dataset based on the recent work of Park et al. [88] on designing a chatbot for a brief motivational interview on stress management. They prepared a total of 220 statements which were reviewed by therapists based on the Motivational Interviewing Skills Code (MISC) [77]. For designing the templated feedback of Stress-bot, we adopted the following approach: each templated response comprised of three short statements; a) An *empathetic response* to the user's problem (e.g., "That must have been very difficult for you."); b) a *high-level reflection* to the user's problem (e.g., "You appear to be frustrated about not being able to have a better work-life balance despite doing your best."); and finally asking a follow-up *open question* (e.g., "What difficulties have you had in relation to your problem?"). On the other hand, a low-quality variant of Stress-bot asks irrelevant questions that are not directly linked to the user's problems (e.g., "That's okay, we can talk about it. By the way what kind of job is it?"). Furthermore, in this condition, Stress-bot did not express empathy, did not reflect or validate the user's concerns and did not ask relevant open-ended questions. Rather, it would wrap up the discussion with an abstract advice that is not totally applicable to the problem (e.g., "Do all those activities which give you the most joy in your life and you will feel better." or "Maybe when you get stressed you can watch something interesting on YouTube.").

Next, based on the quality criteria specified by Xu et al. [123], we constructed two variants of IR-based CPCS in terms of response quality. For instance, a high-quality response is the one that was appropriate and meaningful. For instance, in response to a user query, "how many people in the US are currently living with HIV?", a CPCS with high quality responded with "Thank you for asking. Approximately 1.2 million people in the U.S. are living with HIV today." On the other hand, a CPCS with low-quality response was unsure about its answer and replied with "I think they are slightly over 1 million". For an IR-bot, we crafted answers with insufficient information for low-quality responses and introduced an element of ambiguity to make participants doubt the accuracy of the answer.

3.5 Power and Sample Size Estimation using GPower

The sample size and power for the statistical tests were calculated using the Gpower tool [31, 32]. It is the most widely cited analytical application for commonly used statistical tests in social and behavioral research, and it has been explained and advocated in a number of methodology studies [28, 63]. It supports a range of different power analysis techniques, the most prevalent of which are a priori and post hoc analysis. Priori analyses allow users to determine the sample size necessary to achieve a desired degree of power and effect size. In comparison, post hoc analyses are undertaken following the conclusion of the study and are used to explain non-significant results and estimate test power. GPower has been extensively employed in a variety of fields, including psychology [21, 111], transportation [117], learning and education [74, 79], human-computer interaction research [3, 20], among others.

A priori statistical power analysis using the G*Power application showed that to detect a small effect of $\eta_p^2 = .15$ with 80% power in a 2x2x2x4 factorial between-subjects ANCOVA (32 groups, alpha = .05), we would need at least 489 participants in total. We set the effect size for the current study to 0.15 based on the study [73] that argue that criteria for interpreting experimentally generated effect sizes differ from the frequently cited and extensively utilized guidelines proposed by Cohen. Based on nearly 12,000 correlation coefficients identified in 134 meta-analyses, they advised that correlation coefficients of 0.12, 0.24, and 0.41 and Cohen's ds of 0.15, 0.36, and 0.65 be regarded as small, medium, and large effects in social psychology studies.

We also performed a post hoc power analysis to guarantee that an observed power (or post-hoc power) is sufficient. The sample size of 489 was used for the statistical power analysis with alpha level $p < .05$ and an effect size of 0.15. The recommended effect sizes used for this assessment were as follows: small ($f_2 = .10$), medium ($f_2 = .25$), and large ($f_2 = .40$) (see [22]). The post hoc analysis revealed the statistical power for this study was .80 for detecting a small effect. Thus, our sample size was sufficient to obtain statistical power at the recommended level of .80 [22]. The power to detect a medium-sized effect ($f_2 = .25$) was determined to be 0.99.

3.6 Participants

In total, 522 participants were recruited from Prolific crowdsourcing platform. Out of 522, two submissions were rejected (0.38%), 29 participants returned their submissions (5.56%), and 13 submissions were timed out (2.49%). Hence, our remaining sample included 478 participants. We restricted the experiment to only US and UK participants since our task required proficiency in English. For the stress task, each worker was paid £2.00 fixed amount (£8.00/h), while in the case of information retrieval task, each worker was paid £1.00 fixed amount (£8.57/h). We paid extra money to participants in the stress task because it requires multiple exchanges as opposed to IR task.

The average hourly compensation was £10.55 for the stress task and £9.55 for the IR task, based on the actual amount of time spent by participants. The number of participants from the US was small; there were just 16 for the stress task and 16 for the IR task across all conditions. We also calculated their hourly wage in the US based on the current exchange rate of $1.38 per British pound at the time of authoring this paper. The hourly wage for the stress task was $14.56, while the hourly wage for the IR task was $13.19 (higher than the average minimum wage in the United States of $7.25 per hour). At the time of writing this paper, this hourly wage was categorized as "good" by Prolific's calculator for both US and UK participants. Participants who participated in one condition were not allowed to participate in the other conditions using Prolific's built-in screening feature. The detailed demographics of participants are reported in Table 4.

3.7 Procedure

In this section, we will expand on each stage of the procedure as illustrated in the Figure 4.

3.7.1 Consent. Upon accepting the task participants were asked to read and sign a consent form. This form explained that the chatbot is powered by hybrid intelligence, which means that answers are generated by a combination of computation and human input. More specifically, here are the guidelines we provided to participants regarding CPCS:

> Crowd-powered systems integrate computation with human intelligence, allowing large groups of people to communicate and collaborate online. These "hybrid" systems allow applications to solve problems that people and computation alone

Table 4. Participant Demographics

Characteristics		Stress-bot	IR bot	Total
Gender	Female	165 (69.6%)	160 (66.4%)	325 (67.9%)
	Male	70 (29.5%)	80 (33.2%)	150 (31.4%)
	Prefer not to say	2 (.84%)	0* (0.0%)	2 (.42%)
Nationality	United Kingdom	216 (91.1%)	214 (88.8%)	430 (89.9%)
	United States	21 (8.9%)	27 (11.2%)	48 (10.0%)
Age		33.5±11.8	33.0±11.1	33.3±11.5
Qualification	No Formal Education	0 (0.0%)	7 (2.9%)	7 (1.5%)
	High School Diploma	45 (18.9%)	51 (21.3%)	96 (20.2%)
	College Degree	26 (10.9%)	35 (14.6%)	61 (12.8%)
	Vocational Training	21 (8.9%)	18 (7.5%)	39 (8.2%)
	Bachelor's Degree	94 (39.7%)	91 (38.0%)	185 (38.7%)
	Master's Degree	38 (16.0%)	24 (10.0%)	62 (13.0%)
	Professional Degree	6 (2.5%)	9 (3.8%)	15 (3.2%)
	Doctorate Degree	4 (1.7%)	2 (0.84%)	6 (1.3%)
	Other	3 (1.3%)	2 (.84%)	5 (1.0%)

* We were unable to acquire a participant's gender information because s/he removed his or her response to the gender question before to the study's execution. Due to data privacy laws, Prolific did not provide that information once the participant deletes his or her response.

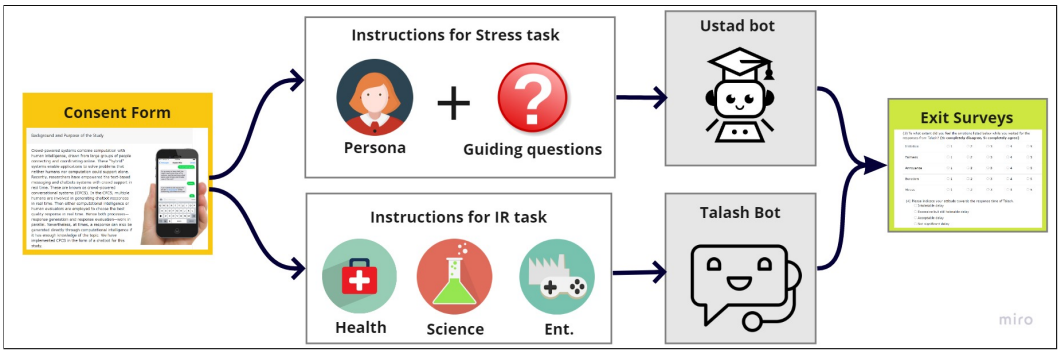

Fig. 4. Stages of Procedure

> couldn't solve. Recently, researchers have added real-time crowd support to text-based messaging and chatbot systems. These are known as crowd-powered conversational systems (CPCS). Multiple humans work together in the CPCS to generate real-time chatbot responses. The best quality response is then chosen in real time using either computational intelligence or human evaluators. Thus, both the response production and response evaluation procedures run concurrently. However, if computational intelligence has sufficient knowledge about the topic, it may sometimes generate a response directly. For this research, we used CPCS in the form of a chatbot. The purpose of the research is to investigate your perception of the chatbot, which is augmented by crowd-powered support in real time.

Participants could then interact with a simulated chatbot that replays human generated statements in a programmed sequence, while controlling the response rate. We notified participants at the end of the experiment that the chatbot's responses were pre-programmed for ethical reasons,

since individuals may exhibit gullibility and embarrassment if they find they have been fooled [34]. Additionally, the funding agency EPSRC [15] developed ethical principles for robotics, highlighting the need of users being able to "lift the curtain" and see how a robot operates on the inside.

3.7.2 Instructions. Participants who read and accepted the consent form were redirected to a detailed instructions page. participants assigned to groups concerning the stress management tasks were presented with the description of a fictional character (see Appendix A for a simple stress task). They were informed that they have to role-play a given fictional character in the conversation. Below the description of the fictional character, we provided some example guiding questions that the crowd-powered conversational system could ask (see Table 5) so that they get in the spirit and anticipate the nature of the choices they would have to make. Because these questions were broad in scope and could be applied to any stress management scenario, we feel they had no influence on the study's naturalistic setting. These generic questions were created using motivational interviewing techniques used by therapists to learn about a stressed person's inner struggles. The purpose of including guiding questions was to acquaint participants with the types of queries the bot might ask. Since the participants' psychological issues were superficial, giving them guiding questions enabled them to play the fictional character as accurately as possible. After they had read about the fictional character and the guiding questions, we let them initiate the conversational task.

In case of IR tasks, participants were presented with three IR tasks (see Table 3) distinguished by the complexity of the conditions in three categories or domains: *Health, Science & Technology and Entertainment.* After they had read about the detailed description of IR search tasks, they could initiate the conversational task.

We did not tell participants that they had to estimate the time in the instructions. We only asked them to estimate the time after they engage with the task, which is called retrospective paradigm [14]. The prospective paradigm, on the other hand, notifies people ahead of time that they need to estimate the time after it has been elapsed. The prospective paradigm is more likely to generate temporal overestimation than the retrospective paradigm [59] because people devote more efforts to judge temporal information and focus less on processing non-temporal information [113]. Therefore, we chose the retrospective paradigm, which is less susceptible to such errors. We do agree, however, that both have benefits and drawbacks. One of the drawbacks of the retrospective paradigm is that people may underestimate the duration because they are unaware that they will be asked to estimate the time and thus they may allocate more resources to non-temporal aspects of the task. Due to these reasons, we used two ways to capture time estimation in the current study: a quantitative estimate of the number of seconds participants thought the bot took, and cognitive appraisal of the wait. Due to the fact that the cognitive construct is evaluative in nature and was deemed to be more substantive in a previous study on time perception [40], we chose it along with the quantitative estimate (seconds) to ensure completeness and to observe the actual time estimates.

3.7.3 Conversational Task. For the stress management task, participants were redirected to the Stress-bot. We adopted Stress-bot based on *complexity, quality*, and *delay* values corresponding to the experimental setup – see Figure 2. The Stress-bot opens a conversation with a predefined utterance: *"Hello, I am Stress-bot, your life coach. I will listen to everything you have to say and help you guide yourself better. Tell me what brought you here today?"*. After that, the user could share their ostensible problems with Stress-bot based on the description of the fictional character. Participants were able to access the description of a fictional character during the conversational task in case they wanted to be reminded of some details (Figure 2). In case of the high quality condition, Stress-bot provides a brief reflection on the user's problem and asks some follow up questions (e.g., *"Could you give me some of the details?"*) to allow the user to expand on her problems

Table 5. Sample Guiding Questions for Stress Task

Sample Questions	Usage
Hi, I would love to hear what that is all about. Do you have any concerns which you would like to talk with me about? Hi, I know you had some things you wanted to talk about, so what brings you in today? I would like to understand how you see things. What brought you here? What has been the problem?	You can expect these questions to be asked at the beginning of the conversation. Explain briefly about your problem (2-3 sentences). You can specify more details later.
What worries you about your problem? How much does that concern you? In what ways? How would you feel about that? Anything specific about your thoughts that made you feel as you do?	These questions will help you to zoom in to your problem further and focus on your most important concerns.
What steps can you take to avoid the frustrations? What have you done to cope with these feelings? What would your partner say about what you are doing?	These questions will help you to explore the options for a change.

further. Finally, after 4-5 dialogue exchanges, Stress-bot provides the user with a piece of advice: *"Great! They say life is about trying things to see if they work. So, let us not give up on all the good things that you are trying, to help cope with this situation! I think you have some positive aspects to act on."* Finally, Stress-bot also asks users about their feelings towards stress (*"How do you feel after talking about it now?"*) with some predefined replies (1: *"I am feeling relieved"*, 2: *"I am feeling neutral"*, 3: *"I am still feeling distressed"*). After selecting an option, participants were informed that this task was completed and they were redirected to the exit surveys.

In the case of the information retrieval task, we deployed an IR-bot across the experimental conditions adapting it regarding the *complexity, quality*, and *delay*. The conversation was initiated by an IR-bot (e.g, *"Hi There! I am an IR-bot, your search assistant. I am so glad you stopped by. Chances are you're looking for something in particular. What can I help you with today?"*), as shown in Figure 3. Following this, users were allowed to select one domain out of the given three domains (*Health, Science & Technology and Entertainment*) – Figure 3 (A). After choosing the domain, the search task was automatically created and displayed to the users in the chat bubble based on the dataset provided by Kelly et al. [57] – Figure 3 (B). After a set delay, participants were shown a templated response. Finally, users were presented some quick replies (*"OK Thank You!"* or *"OK Perfect!"*) which they needed to select in order to exit the IR task – Figure 3 (C). We did not provide users an option to repeat the IR tasks or allow them to ask follow-up clarification questions, so as to control the complexity of the search task and the quality of the responses by an IR-bot.

We confined an IR-bot to only one turn in comparison to the Stress-bot due to the practical distinction between the bots' natures. A task-oriented dialog system aids the user in completing specific tasks within a domain, for example, restaurant bookings, weather searches, and simple information retrieval. Typically, the emphasis of the conversation is brief and limited, and the purpose is usually reached with the fewest possible turns [42, 126]. In comparison, the primary goal of emotion-driven discussion is to enhance user engagement and rapport [2, 101, 106]; as a result, the conversation involves various turns and is typically broad and deep in scope [126]. Thus,

it was natural for Stress-bot's participants to spend more time than IR-bot's participants. However, we did not impose a time limit on participants in either case.

After the conversational task, participants were asked to fill out the exit surveys. In the next section, we expand upon the exit surveys and measures.

3.8 Measures & hypotheses

3.8.1 Perceived Waiting Time. The perceived waiting time was assessed by asking the participants to give an estimate of the total time (in seconds) they had spent waiting between responding to the system and receiving the bot's response. More specifically, we asked: *"How did you perceive the response time from Stress-bot/IR-bot to be during your conversation? Indicate your estimate in seconds by simply entering the number of seconds"*

3.8.2 Cognitive and Affective Appraisals of the Wait. Attitude is considered as comprising of two distinct dimensions: cognitive appraisals and affective appraisal [58]. Cognitive appraisals represent the utilitarian aspect of the attitude, which is used for the products that have functional value; for instance, a grammar checker software may be evaluated cognitively [116]. We measured cognitive appraisal based on the perception of time spent in terms of long or short judgement [93]. It is measured on a five-point scale (1: 'very short' , 5: 'very long'). In the survey, we asked participants: *"How would you best describe the response time from Stress-bot(IR-bot) during your conversation?"*

On the other hand, affective appraisals refer to the hedonic aspect of the attitude, which is aimed for self-fulfilling rather than functional value; for example, the evaluation of the computer game is based on the affect [116]. More specifically, affective appraisals of the wait capture the emotional response of users towards the response delays. This scale is sufficiently reliable (Cronbach alpha is 0.80) and one-dimensional [93]. It consists of five semantic items asking participants to rate the extent to which they have experienced irritation, unfairness, annoyance, boredom, and stress while waiting. Each item is measured on a five-point scale (1: "completely disagree" and 5: "completely agree"). In a survey, we explicitly asked participants: *"To what extent did you feel the emotions listed below while you waited for the responses from Stress-bot/IR-bot?"*. We compute the mean affective experience score by averaging these items. Table 6 lists hypotheses regarding the cognitive appraisal of the wait and perceived quality.

3.8.3 Perceived Response Quality. We also measured the perceived response quality of Stress-bot and IR-bot on a 10-point scale (1: "Very low quality", 10: "Very high quality"). Specifically, we asked them: *"How did you perceive the response quality provided by Stress-bot (IR-bot)"*.

3.8.4 Qualitative Feedback. We also asked participants to share any additional thoughts, remarks, or feedback that they may have regarding their experience interacting with Stress-bot (IR-bot). In addition to that, we also asked participants to provide their suggestions on handling long delays in CPCS.

3.8.5 Control variables. In this section, we describe individual factors that can mediate the effects hypothesized above. Thus, we treated them as covariates or control variables in the statistical model.

(a) Experience with Chatbots: Recent work has shown that experience with chatbot technology can impact the users' self-reported preferences. For instance, a recent study shows that users with more experience with chatbot technology desired more detailed features that can explain the decisions made by the underlying machine learning models [9]. Another study shows that more experienced users are likely to express more critical opinions about the chatbot [71]. Thus we measured the self-reported prior experience of participants with chatbots using two-item scale (*"I am familiar with chatbot technologies"* and *"I use chatbots frequently"*). We measured these items

Table 6. Hypotheses regarding appraisals of the wait and perceived quality

RQs	Hypotheses	References in support of the hypothesis
RQ1	H1: The longer the response time, the more negative the appraisals of the wait (both cognitive and affective)	[18, 104]
RQ2	H2: A poor-quality response of CPCS leads to a relatively more negative appraisal of the wait in comparison to a high-quality response. H5: The higher the quality of the bot's responses, the greater the perceived response quality.	[30, 93, 125]
RQ3	H3: The higher the complexity of the task in a CPCS, the more positive are the appraisals of the wait.	[5, 13]
RQ4	H4: Longer response times leads to relatively more negative appraisals of the wait in the stress mitigation task than the information retrieval task.	[89, 107]

on 7-point scales (1: "Completely disagree", 7: "completely agree") [9]. This scale showed good reliability (Cronbach alpha=0.71).

(b) Social Orientation Towards Chatbots: The social orientation scale [71] measures the preference of people towards realizing chatbots as "social agents". People who have a greater tendency to treat chatbots as social entities are more likely to engage with them through natural conversation compared to those who have low tendency. Thus, it is possible that people who treat chatbots as social entities would evaluate the waiting experience more positively. The social orientation towards chatbots was measured by using a two item scale (*"I think "small talks" with an AI agent or chatbot is enjoyable."* and *"I like chatting casually with a chatbot."*). We measured these items on 7-point scales (1: "Completely disagree", 2: "completely agree"). This scale has good reliability (Cronbach alpha=0.86).

(c) Affinity for Technology Interaction This scale can measure the users' general technological affinity to actively engage with the technology. We were also interested to investigate whether people who have a greater inclination towards technology (e.g., *"I like to occupy myself in greater detail with technical systems."*) could evaluate the waiting experience differently than those who do not have any inclination or were technophobic (e.g., *"It is enough for me that a technical system works; I don't care how or why."*). The affinity towards technology interaction (ATI) scale [33] is a 9-item scale where each item is measured on a 6-point scale (1: "completely disagree", 6: "completely agree"). This scale has an excellent reliability (Cronbach alpha=0.88 to 0.92) supported by multiple studies with over 1500 participants. [33].

3.9 Data Analysis

We planned to study the effects of *context*, *complexity*, *quality*, and *delay* on the *perceived waiting time*, *cognitive (Long/short judgment)*, *affective waiting experience*, and *perceived response quality* using 2x2x2x4 factorial ANCOVA by treating social orientation towards chatbots, experience with chatbots and affinity for technology interaction as covariates. There was no strong correlation between the covariates, as assessed by multicollinearity test ($r < 0.3$). We also performed the homogeneity of regression slopes test before conducting ANCOVAs by investigating the interaction between highest order interaction and covariates; the assumption of homogeneity of regression

slopes was met for all outcome variables ($p > .005$). In a few cases, the assumptions of normality and/ or homogeneity or variances were not met. However, ANCOVA is robust with respect to deviations from homogeneity of variances and deviations from normality [10]. We removed five cases from the dataset because participants provided anomalous feedback. Except in three cases (two conditions had 13 participants and one had 14), each condition had fifteen participants in total. This happened because Prolific's built-in algorithm immediately employs more participants to compensate for those who left the task. Due to the fact that we had just three cases in which the number of participants was slightly fewer than the needed number, we anticipate that heterogeneity of variance will not be an issue, as ANOVA is considered robust to modest deviations from this assumption [41].

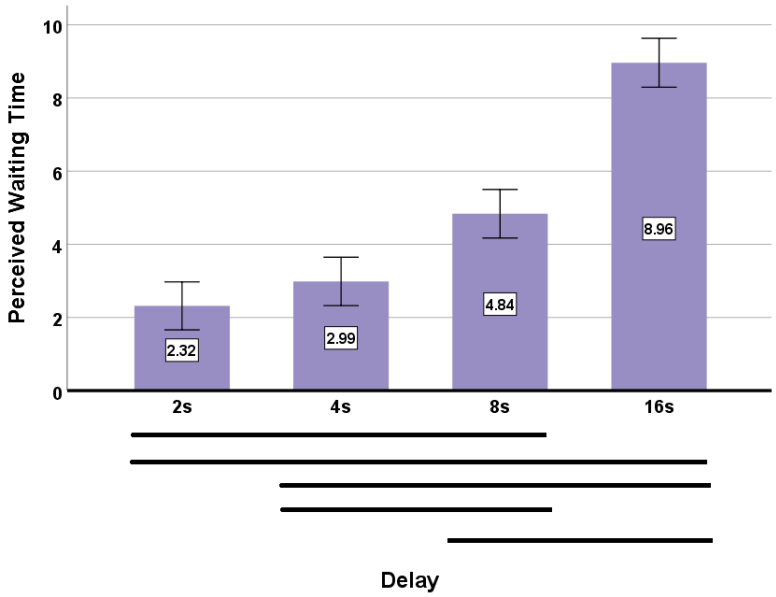

Fig. 5. Perceived waiting times means and standard errors for the four levels of the actual delay (independent variable). The analysis showed a main effect of delay on mean perceived waiting time. Statistically significant differences in pairwse comparisons are represented with the line segments under the horizontal axis

4 RESULTS

4.1 Perceived Waiting Time

Before conducting an analysis, we deleted the data instances based on the three standard deviation criteria. Factorial ANCOVA revealed no interaction effects (4-way, 3-way, and 2-way) between all four factors for the mean perceived waiting time. However, the simple main effect of delay on mean perceived waiting time was statistically significant ($F(3, 427) = 77.90$, $p < .001$, $\eta_p^2 = 0.354$ – (Fig. 5). Pairwise comparisons were made for all levels of delay with a Bonferroni correction. The mean perceived waiting time score was 2.52 ($95\%CI$, 1.26 to 3.78) points higher for 8s delay than 2s delay ($p < .001$), 6.65 ($95\%CI$, 5.38 to 7.91) points higher for 16s delay than 2s delay ($p < .001$), 1.85 ($95\%CI$, .582 to 3.11) points higher for 8s delay than 4s delay ($p = .001$), 5.98 ($95\%CI$, 4.70 to 7.25) points higher for 16s delay than 4s delay ($p < .001$) and 4.13 ($95\%CI$, 2.86 to 5.40) points higher for 16s delay than 8s delay ($p < .001$).

4.2 Cognitive (Long/short Judgment)

Factorial ANCOVA showed a statistically significant main effect of context (stress, IR) on the cognitive score, $F(1, 441) = 4.34$, $p = .038$, $\eta_p^2 = 0.010$). This supports our hypothesis (**H4**) that participants in the stress task judged response delays to be slightly longer ($M = 2.09$ ($95\%CI$, 1.98 to 2.20)) than those who interacted with an IR-bot in the information retrieval tasks ($M = 1.92$ ($95\%CI$, 1.81 to 2.03))–see Figure 6.A. Note that high scores given to the cognitive delay represents that the response delay was perceived to be longer.

We also found a statistically significant main effect of complexity (low, high) on cognitive score, $F(1, 441) = 5.73$, $p = .017$, $\eta_p^2 = 0.013$). This supports our hypothesis (**H3**) that when the complexity of the task was lower, participants perceived response delays to be slightly longer ($M = 2.10$ ($95\%CI$, 1.99 to 2.21)) than the high complexity task ($M = 1.91$ ($95\%CI$, 1.80 to 2.02)) – see Figure 6.B.

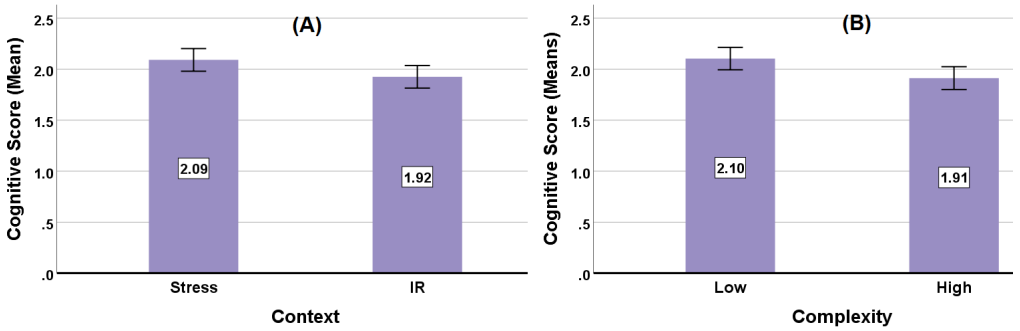

Fig. 6. **(A)** Result of simple main effect of *context* (stress, IR) on the mean cognitive score. Results indicate that the participants in the stress tasks perceived response delays to be slightly longer than IR tasks; **(B)** Simple main effect of *complexity* (low, high) on the mean cognitive score. Participants perceived response delays longer when the complexity of the task was lower

The simple main effect of delay on mean cognitive score was statistically significant ($F(3, 441) = 44.96$, $p < .001$, $\eta_p^2 = 0.234$) – Figure 7. The mean cognitive score was 0.53 ($95\%CI$, 0.23 to 0.81) points higher for 8s delay than 2s delay ($p < .001$), 1.15 ($95\%CI$, 0.85 to 1.44) points higher for 16s delay than 2s delay ($p < .001$), 0.48 ($95\%CI$, 0.18 to 0.78) points higher for 8s delay than 4s delay ($p < .001$), 1.10 ($95\%CI$, 0.80 to 1.40) points higher for 16s delay than 4s delay ($p < .001$) and 0.62 ($95\%CI$, 0.32 to 0.92) points higher for 16s delay than 8s delay ($p < .001$) – Figure 7. This supports our hypothesis (**H1**) that when the response time was longer, participants evaluated the cognitive appraisals of the wait more negatively.

4.3 Affective Waiting Experience

Prior to presenting results, it is necessary to note that a high score of affective represents negative responses (annoying, irritating, boring, unfair, stressful). Through factorial ANCOVA, we found a statistically significant main effect of *quality* (low, high) on *affective* score, $F(1, 441) = 17.00$, $p < .001$, $\eta_p^2 = 0.037$). This shows that when the quality of the bot's responses was lower, participants experienced more negative affective emotions ($M = 2.06$ ($95\%CI$, 1.97 to 2.16)) than in case of high quality bot responses ($M = 1.78$ ($95\%CI$, 1.68 to 1.87)) – Figure 8. This supports our hypothesis **H2**.

Furthermore, the simple main effect of *delay* on mean *affective* score was statistically significant ($F(3, 441) = 14.51$, $p < .001$, $\eta_p^2 = 0.090$) –Figure 9. Pairwise comparisons were made with a Bonferroni correction. The mean affective score was 0.55 ($95\%CI$, 0.287 to 0.805) points higher

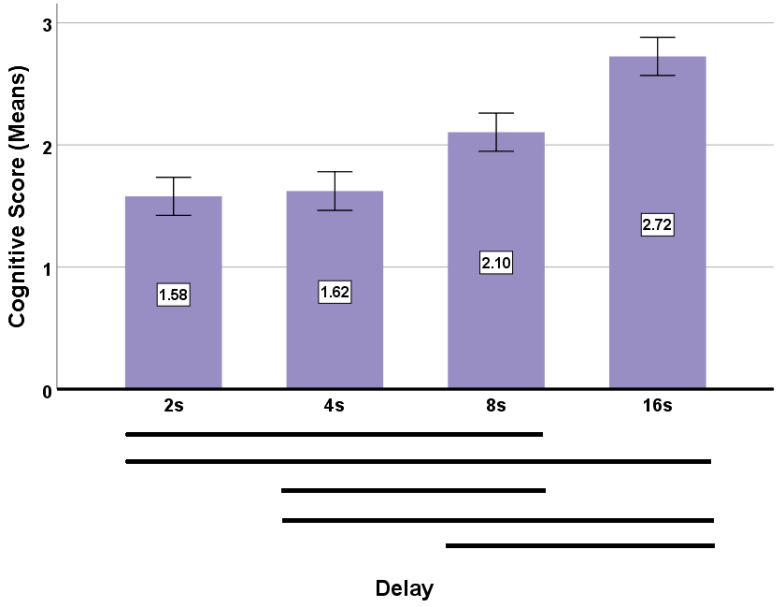

Fig. 7. Mean cognitive scores and standard errors for the four levels of the actual delay (independent variable). The analysis showed a main effect of delay on the mean cognitive score. Differences found significant with pairwse comparisons are represented with the line segments under the horizontal axis

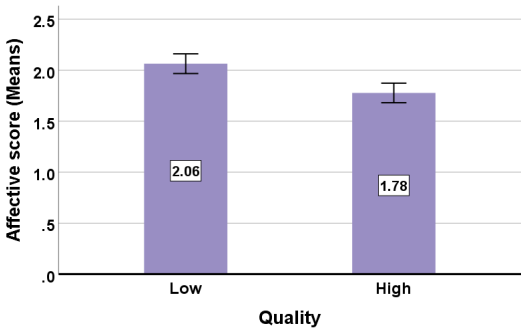

Fig. 8. The simple main effects of *quality* on the mean *affective* score. Results show that when the quality of the bot's responses was lower, participants experienced more negative affective emotions

for 16s delay than 2s delay ($p < .001$), .53 ($95\%CI$, 0.266 to 0.788) points higher for 16s delay than 4s delay ($p < .001$) and 0.50 ($95\%CI$, 0.244 to 0.762) points higher for 16s delay than 8s delay ($p < .001$). This shows that participants experienced more negative affective waiting experience when they interacted with the bot that responded to them in 16s (Hypothesis **H1**). There was no significant difference in the affective waiting experience scores between 2, 4 and 8s delays.

We also found a significant main effect of the covariate *social orientation towards chatbots* on the mean *affective waiting experience* scores: ($F(1, 441) = 7.09$, $p = .008$, $\eta_p^2 = 0.016$). A negative correlation between *social orientation towards chatbots* and *affective waiting experience* was observed

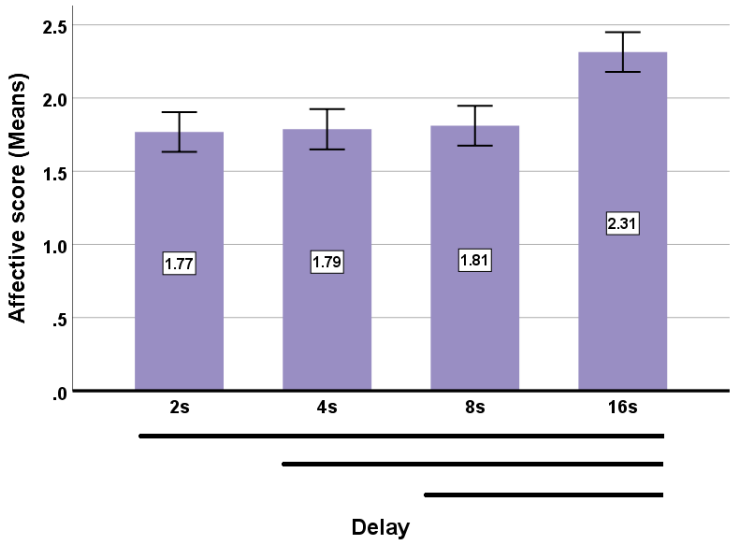

Fig. 9. Mean affective scores and standard errors for the four levels of the actual delay (independent variable). The analysis showed a main effect of delay on the mean affective score. Differences found significant with pairwse comparisons are represented with the line segments under the horizontal axis

($r = -.15, p = .001$), which means that participants who perceived chatbots as more social entities rather than machines perceived less negative affective emotions.

It should be emphasized that, while we employed two distinct task types, each of which has the ability to vary the complexity of the task, we make no comparisons of task complexity"between" the two task types. For example, we discovered that when workers were assigned to "high-complexity" tasks using the Ustad bot, their emotions were severely impacted compared to when they were assigned to IR tasks using the Talash bot. However, since each task's complexity is defined differently, the comparisons are unfair. For example, in the case of the IR task, high complexity task was determined using the cognitive complexity framework, which defines a more complicated scenario as the one that demands extensive problem-solving skills, such as restructuring pieces into a new pattern or structure by creating, planning, or producing. While in the instance of stress management, a more complex task was one involving many stressors, which necessitates a different set of supporting conversational abilities (demonstrating reflective listening, expressing empathy, asking focusing or evoking open questions) than those required for IR tasks. Therefore, even though we found a significant interaction effect between task type and complexity on the affective score, we did not report it. We incorporate two levels of complexity into each task type, inspired by the seminal work on task complexity [122], and then compare how this may affect time perceptions. In other words, we compare complexity "within" each domain separately, namely within the IR and stress mitigation domains. We investigated task comparisons in a separate research question (RQ4), for which we constructed two distinct chatbots with distinct roles and purposes.

4.4 Perceived Response Quality

Factorial ANCOVA revealed a significant interaction effect between *complexity* and *quality* ($F(1, 441) = 10.34$, $p = 0.001, \eta_p^2 = 0.023$). The simple effect test shows that there was a statistically significant simple effect of *complexity* on the mean perceived response quality scores when the

quality of the bot responses was lower ($F(1, 232) = 30.15$, $p < .001$, $\eta_p^2 = 0.115$), but not for high quality bot responses ($F(1, 234) = 2.51$, $p = .115$, $\eta_p^2 = 0.011$) – Figure 10.A. Overall, when the quality of the responses was lower, participants in the low complexity tasks perceived the quality of interaction with the bots more positive ($M = 6.09$ ($95\%CI$ 5.68 to 6.50)) than when the complexity of the tasks was higher ($M = 4.45$ ($95\%CI$ 4.03 to 4.86)) – Figure 10.A. Nevertheless, complexity did not influence when the quality of the bot's responses were high. These results show that hypothesis **H5** does not hold when the complexity of the task is higher.

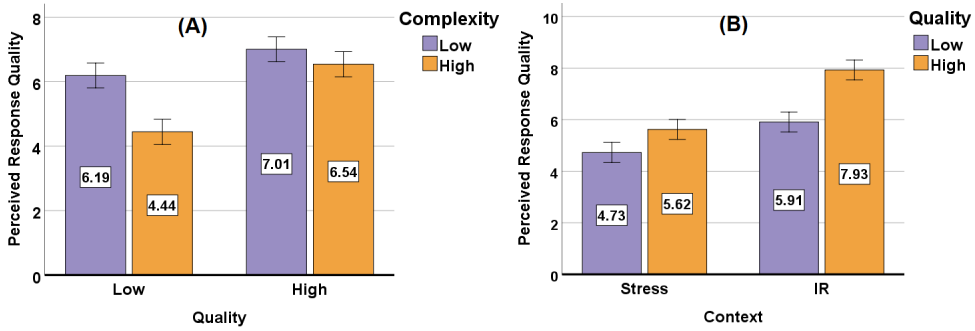

Fig. 10. **(A)** Interaction effects between *Quality* and *Complexity*: Results indicate that when the quality of the responses were lower, participants in the low complexity tasks perceived the quality of interaction with the bots more positive than when the complexity of the tasks were higher; **(B)** Interaction effects between *Quality* and *Context*: Result shows that in both contexts (stress and IR), participants significantly differentiated between the high and low quality responses.

We also found a significant interaction effect between *context* and *quality* ($F(1, 441) = 8.21$, $p = .004$, $\eta_p^2 = 0.018$). The simple effect test shows that there was a statistically significant simple effect of quality on the mean perceived response quality scores when the context was stress ($F(1, 232) = 9.44$, $p = .002$, $\eta_p^2 = 0.039$), and also when the context was IR ($F(1, 234) = 45.09$, $p < .001$, $\eta_p^2 = 0.162$). This shows that in both contexts (stress and IR), participants significantly differentiated between the high and low quality responses (supporting **H5**). Overall, participants perceived the interaction quality with an IR-bot more positive than the Stress-bot (Figure 10.B).

We also found significant main effect of the covariate *social orientation* on the mean *perceived response quality* scores: ($F(1, 441) = 49.93$, $p < .001$, $\eta_p^2 = 0.102$). A positive correlation was found between *social orientation* and *perceived response quality* ($r = .345$, $p < .001$) which means that participants who perceived chatbots as social entities rather than machines perceived the response quality to be higher. We have provided the summary of key findings in the Table 7.

4.5 Suggestions on Handling Long Delays

To gain insights of participants' opinion about the design of CPCS for handling long delays, we calculated the number of suggestions that participants described and presented them in a Figure 11. More specifically, we asked participants: *"Do you have any suggestions for how conversational systems can handle long delays in responses?"*. We received 477 feedbacks in total from participants. The collected feedback responses were subjected to an inductive thematic analysis [17] using the Dedoose program[7]. The first iteration assessed responses and generated initial codes. The second iteration analyzed previous themes and integrated those that were redundant. After two coding

[7]https://www.dedoose.com/

Table 7. Summary of the key findings with respect to our research questions and hypotheses.

RQs	Hyp.	Supported?	Summary of key findings	Reference
RQ#1	H1	Yes	When the response time was longer, participants evaluated appraisals of the wait more negatively.	Section 4.2, 4.3
RQ#2	H2	Partly	• When the quality of the bot's utterances was lower, participants experienced more negative affective emotions. • We did not find support for the cognitive component (long/short judgment).	Section 4.3
	H5	Partly	• Participants in both contexts (stress, IR) significantly differentiated between high and low quality responses. • When the quality of the bot's utterances was lower, participants perceived the quality of interaction with the bots more positive *only when the complexity of the task was lower.*	Section 4.4
RQ#3	H3	Partly	• When the complexity of the task was lower, participants perceived response delays to be slightly longer. • When the complexity of the task was higher, participants experienced more negative emotions.	Section 4.2, 4.3
RQ#4	H4	Yes	Participants in the stress task judged response delays to be slightly longer and experienced more negative emotions than those who were in the IR task.	Section 4.2, 4.3

cycles, the analysis had settled on the primary issues discussed in this paper. Now we briefly expand on each suggestion in the next sections.

4.5.1 Holding or Filler Messages. The majority of the participants recommended filler or holding messages (N = 98). Conversational fillers or holding messages are quite ubiquitous in human conversations, such as "well", "hm", "uh" or "let me think" [29]. They let the waiting person know that s/he is not ignored and the response is likely to come shortly. For instance, One worker (ID:385) replied : *" Perhaps they can provide a quick response first such as 'I understand, please give me a moment to think on what you have said' before issuing a more in-depth response... "*. Another worker (ID:370) replied: "Maybe a holding message to say, 'please bear with me while I think about that'?"

4.5.2 Typing indicators. Typing indicators make chatbot interactions appear more natural. These indicators show that a message is being typed and thus could encourage users to wait. participants mentioned two types of typing indicators that are very prevalent in messenger applications; 1) graphical typing indicators (N = 38); 2) text-based typing indicators (N = 11). Graphical typing indicators include three animated dots [...], progress bar or loading or thinking icon while the text-based indicators include "Person X is typing". For example, concerning the graphical indicator, one worker (ID:264) replied as *"Maybe add ... to show that the chat bot is currently typing. There was nothing to illustrate to me that the chat bot was actually working on its next response to me which was a little annoying."*. Example of text-based indicator was described by worker (ID:271) as follows:

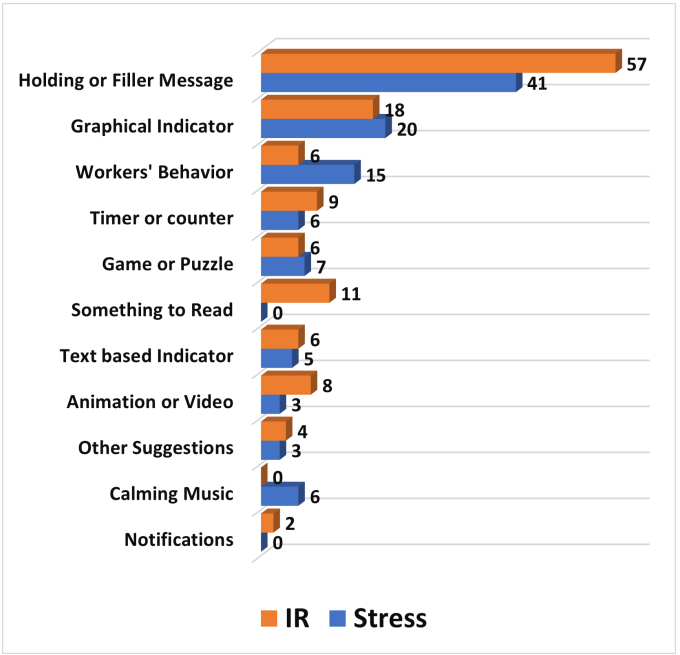

Fig. 11. participants' suggestions about handling the long delays in crowd-powered conversational agent

"I think having a notification saying 'x is typing' makes the waiting process more realistic and more reassuring"

4.5.3 Using Crowd's behaviour. Participants also suggested solutions to reducing response delays that were directly linked with the crowd's behaviour (N = 16), such as checking whether other person (worker) is still there, communicating with workers beforehand about the expected response delays, having a 'seen' function akin to messaging applications, send reminders to workers who are slow in responding. They also introduced a strategy to show partial responses that allow the user to see a response when it is being written in real time. This is similar to the IMO messaging app that allows you to see in real-time what the other person is typing. For instance, *"Check that the other person is still there..."*(ID:394) or *"Communicate with the person how long response times are expected to take at the beginning of the conversation."*(ID:278) . Regarding the partial responses, one worker (ID:268) said, *"Just to display the word typing or similar so you know you will definitely get a response; the waiting was frustrating."*

4.5.4 Entertaining Activities. Participants also mentioned entertaining activities as a way to distract a waiting person. This includes playing some game or puzzle (N = 13), something interesting to read (N = 11), watching a cartoon animation or video (N = 11) and listening to some relaxing music (N = 6). [Game or Puzzle, ID:253]: *"There could be simple games like 'Pong' or 'Snake' that can keep the user occupied while waiting"* [Short readings, ID:66]: *"...Giving the user something short to read may make the delay seem less lengthy."*. [Animation or Video, ID:461]: *"Perhaps have an interesting character moving around the screen"*.

4.5.5 Notifications. Participants also mentioned that when the delays are longer, one's attention may wander off to some other tasks. Therefore an audio notification at short interval may help to

retain one's attention (N = 2). Example includes (ID:63): *"Unsure as sometimes delays are unavoidable. It would be good to have a ding sound when the bot responds so you can click away on other tabs, then be alerted when there is a response"*

4.5.6 Other suggestions. Other suggestions include showing emojis, engaging users with some breathing activities and providing tools for note-taking or thought recording while waiting.

4.6 Participants' Perceptions about Different Delay Levels

Using the same thematic analysis approach as outlined previously, we investigated instances in which participants expressed their emotions concerning varying levels of delay. Overall, participants from 2s (N = 38) and 4s (N = 44) conditions believed that the responses in both conditions were quick because it is impractical for humans to provide in-depth answers in such a short span of time. For instance, a participant in 2s condition responded as [ID.35]: *"The response is very fast, so I know that this is not a human typing a response. The response is quite meaningful and thorough. I am happy with the response."* Another participant in 4s condition replied [ID.186]: *"The responses from Stress-bot were very fast so sometimes it felt like he hadn't really understood but after reading the responses I found they were appropriate"*.

Participants in the 8s (N = 25) condition believed that the response time was not too long. This assertion was echoed by a participant [ID.193]: *"The response time was good, and I was not kept waiting a long time. If this happened on all chatboxes I don't see how anyone would complain"*. Another participant [ID. 394] who interacted with Stress-bot commented: *"I don't feel the delay in responses was excessive. As a patient, if I were waiting for that level of response from a human, I'd be happy they thought about it…"*.

Only a few participants (N = 13) supplied feedback on the 16s condition; those who did provided feedback were equivocal about their feelings. Some participants (N = 5) thought the response time was irritating, while the remainder (N = 8) thought it was reasonable. For instance, one participant [ID.456] expressed dissatisfaction with the long delay as follows: *"I thought the response was well-thought out, informative and interesting, but the delay made me think I did something wrong, or something was broken at first"*. Another participant [ID. 156] replied: *"Even though I expected the delay, the thought did flash across my mind that perhaps I wasn't going to get a response and my internet had frozen, when I was sitting waiting…"*. Although delayed responses negatively impacted on the user experience with CPCSs, it increased their perception that a real human was talking instead of a bot; one participant [ID.216] replied: *"It felt like an actual person and not a chatbot"*. One participant [ID.460] who believed that the response was reasonable commented: *"IR-bot gave a good response and answered my question well within a reasonable amount of time."*. Another participant [ID.517] responded as: *"no I think the response time was fine, I have waited longer for responses in the past."*

5 DISCUSSION AND CONCLUSIONS

5.1 RQ1: 2-4s Delay is Robotic, 8s Delay is Tolerable but 16s Delay is Captive

Our findings indicate that participants perceived the waiting time to be less than it actually was and were unable to differentiate between two and four second delays (cf. Figure 5). Thus, it is obvious that the functional or utilitarian appraisal of time is valid until four seconds of delay. This is owing to the engaging nature of digital interventions, particularly chatbots, which mimic human connection through a conversational approach. This might have helped participants to divert their focus away from the time until 4s and increased their engagement [91, 92, 95]. Theoretical models developed by psychologists concerning time perception models can also explain this phenomenon. For example, cognitive absorption [4] is defined as a 'state of intense involvement with software'

(p. 673) and is derived from Csikszentmihalyi's flow theory [24], which defines a mental state in which a person performing an activity is completely immersed in a sense of energized focus, deep involvement, and enjoyment in the process of the activity. When the system took longer than 8s, however, the emotional component was affected. Given that this deep involvement is linked to subjective experiences with cognitive and emotional dimensions [91], it is likely that negative experiences concerning the appraisal of the wait started to slowly emerge after 4s but worsened after 8s. In the 2 and 4s conditions, the uncanny valley theory [81] may account for participants' assessment of the bot's utterance as robotic and unexpectedly fast. For instance, a study shows that participants perceived interaction with the chatbot as an unpleasant or eerie if they have to wait longer for the response especially when chatting with a human [110]. In our case, since the participants knew a priori that real humans are generating responses, they might have assumed that they would take some time in composing a valid response. Thus, an extremely fast response made them uncertain regarding the true nature of a CPCS as human or AI that provoked feelings of eeriness. This finding also reflects that the development of technological innovation to lower the response latency below 4s would be extravagant within the context of CPCS.

Participants from the 8s condition deemed that the response time was not excessively long. This finding seems noteworthy but poses an important question: Is it conceivable to attain up to an 8 second response latency for difficult conversational tasks in CPCS that require more cognitive capacity? One solution is to use speech to text (STT) services on the crowd interface. For instance, CoZ [2] was able to endow workers' web interface with STT API for quickly transliterating voices of workers into text. Furthermore, people can speak faster (compared to typing) and take shorter pauses– typically about a quarter to half a second [86]–to reflect on what they say. Therefore, an 8 second duration seems sufficient for people to come up with a reasonable answer with STT. These results are also consistent with a prior finding about system response time in telephony speech applications where authors claimed that waiting time should not exceed beyond 4–8 seconds [35]. Favoring this claim, Nielsen [85] suggests that 10 seconds is the upper limit for keeping the user's attention focused on the dialogue.

The 16s condition elicited higher negative affective reactions from participants (cf. Figure 9). Similar to our finding, Yang et al. [125] showed that delays in the response of conversational agents were associated with negative affect (distressed, anxious, tense etc.). Likewise, Butler [18] used the same geometric sequence of response delays in simple computer-based data entry tasks. He found that when the computer took longer time to respond in printing the prompt characters, so did users in responding. Furthermore, Chatbots waits are usually open-ended, where no explicit waiting duration can be conveyed after a user inquiry. Thus, a user's attention is held captive until a chatbot response [121]. Miller [76] wrote in his report about response time in conversational transactions that "captivity of more than 15 seconds, even for information essential to him, may be more than an annoyance and disruption".

In summary, within the CPCS context, researchers need to develop algorithms and workflows to keep the response latency within 4-8 seconds. For scenarios that require longer response time, such as complex information retrieval tasks, the effects of extended waiting time can be alleviated by designing and testing different time fillers [121]. Besides, the fact is that current CPCS are nowhere near the threshold of 8s; for instance, field deployments of Chorus show that 25% percent of the conversations obtained a first answer within 30 seconds [46]. Although CoZ [2] was able to keep response latency under ~8 seconds by broadcasting an audio-video stream, this is not an affordable solution for long-term deployments. Thus, our current study is the first that addresses the issue of time perceptions in CPCS and raises the urgency of developing techniques beyond already known techniques (e.g. pre-recruiting and retainer [11, 45]) to reduce the perceived latency of CPCS within the desirable range of 4-8 seconds.

5.2 RQ2: Quality of Response's Content can Influence Waiting Experience

Participants perceived the affective response in terms of irritation, boredom, fairness, etc. more negatively when the quality of the bot's responses was lower, regardless of response delays (cf. Figure 8). These findings are in line with the prior work where authors studied the impact of product qualities (measured with pragmatic quality, hedonic quality, and the overall appeal) on the affective experience with conversational agents [125]. One important finding was that "quality of response content" was negatively related to the negative affect, that is, when the quality of the response content was lower, participants experienced more stress.

Given that one of the criteria for response quality in the current study was usefulness, we also analyzed the association between response times and perceived quality (Cf. section 4.4). Our findings, however, did not provide solid evidence regarding the influence of the bot's response quality on perceived delay, other than the fact that good quality responses reduce frustration. However, it appears that when the quality of the bot's answers was high, the estimation of passing time would be slightly compressed when the delays were longer; 8.8% decrease for 8s delay and 4.7% decrease for 16s delay. Thus, the high-quality of responses can be an important factor for countering the negative effects of waiting. However, further studies within the context of CPCS are needed to test this association. Additionally, because cognitive appraisal of the wait is related to time, there is ample evidence that shows that shorter delays result in greater satisfaction (see [51, 94, 112]). In other words, the more positive the wait is rated, the more satisfied the user is with the service.

To the best of our knowledge, no study in human-chatbot interaction or in CPCS has looked specifically at the influence of quality of response content on the perception of time. We are only aware of few studies in the service marketing and transit agency context that study the relationship between quality of service and perception of waiting time. For instance, Pruyn and Smidts [93] studied the attractiveness of the waiting environment in terms of comfort, spaciousness, and atmosphere on the perception of waiting time. They argued that quality of environment can serve as an element of distraction and consequently can lower the perception of elapsed time, though their study did not reveal such effects. The reason we found only a weak relationship between the response quality and perception of time could be for a variety of underlying factors; for instance, it may be the case that variation in the quality of the bot's responses in this study was not substantial enough to evoke the potential effect on the perceived waiting time.

These results also provide an interesting trade-off between content quality and latency. For instance, researchers have developed and tested a plethora of techniques to lower the objective waiting time in CPCS through pre-recruiting, queuing, and other computational techniques [11, 43, 45]. Nevertheless, lowering the objective waiting time beyond certain limit could compromise the quality of crowd output due to time pressure on workers. Our result seems to support this notion that the quality of response contents is an important factor that may influence the perception of time passing and thus, by uplifting the quality of response contents one can lower the perceptions of actual waiting time. Maintaining the high quality of crowd output is an active area of research in crowdsourcing [61]; for instance, research has shown that properly training crowd workers to support complex conversational tasks could significantly improve the quality of response content [3]. Our research has advanced the state of the art in CPCS by stating that subjective waiting time perceptions can be improved by enriching the quality of response content. As a result, this can ease the engineering efforts in real-time crowdsourcing research to improve the response latency to a desirable extent of 8 seconds based on our results.

5.3 RQ3: Perceived Time Drags and Perceived Quality Boosts Up with Low Complexity Tasks

When the tasks' complexity was lower, participants overestimated the elapsed time in comparison to high complexity tasks (Figure 6.B). Past research shows that a complex activity can lead to underestimation of elapsed time [5, 13, 59]. For instance, this assertion was supported by Khan et al. [59] who conducted an experiment to find the effect of cognitive load on time perception. The complexity of task was varied based on the cognitive load—ranging from merely paying attention to memorizing the presented items. Their results revealed a negative association between cognitive load and time perceptions, that is, when the cognitive load increased, participants underestimated the time perceptions. Likewise, Block and Gellersen [13] varied the complexity of three input techniques, including simple on-screen icon invoked by the mouse (GUI), graphical toolbar that was mounted on the keyboard to invoke commands (display keyboard) and finally a touch-sensing keyboard (touch keyboard) combined on-screen icon with corresponding touch-based keys on the keyboard. In their experiment, a touch keyboard was considered to be a cognitively more stimulating input modality than others. Their results indicate that participants who executed commands with the touch keyboard perceived the time to be shorter.

This inverse relationship between cognitive load and time perceptions can be better explained with the well-grounded *cognitive-timer* [38] and *cognitive-attentional* [113] models from psychology. For instance, the cognitive–attentional [113] model explains that a person's attentional resources are limited and are usually split between temporal and non-temporal information during an activity. Thus, if an activity has a non-temporal dimension, such as recalling more information, then the temporal estimation is affected. The cognitive-timer models also relies on similar assumptions: *"(1) the existence of a cognitive timer whose purpose is to process and generate temporal information; (2) temporal information is processed by the timer by storing the number of subjective time units which have accumulated during a given interval; and (3) attentional resources are allocated continuously to enable both temporal and non-temporal information processing, with a trade-off in allocation between the cognitive timer and other cognitive modules"*. For instance, a high complexity search task involves more cognitive modules, such as extracting knowledge from different sources and then structuring and organizing them to create a new idea, which results in a decreased attention towards the temporal aspects; Thus, it is possible that the user may establish a belief around this notion that CPCS may take longer to respond and consequently their biological mechanism—the internal clock–[64] is adapted to underestimate the temporal judgment. On the other hand, a low complexity search task only requires extracting a relevant fact from the information sources, which requires less cognitive resources and attention is devoted towards temporal aspects. Therefore, the time appears to pass more slowly in the low complexity tasks. Thus, if CPCS takes longer to respond in low complexity tasks, it can further exacerbate users' waiting experience.

We also found a combined impact of task complexity and response quality on the perceived interaction quality with CPCS. When task complexity and response quality was low, participants perceived the interaction quality higher compared to when task complexity was high (Figure 10.A). Results obtained by Borromeo et al. [16] are consistent with our findings who studied the impact of task complexity and crowd type on work quality. They tasked workers with two types of tasks; a complex task required workers to extract more information items (e.g., title, authors, source etc.) from a given presentation, whereas a simple task required workers to only extract one information item. Their experiments indicate that the simple task generated higher quality results than the complex task by a paid crowd. Another possible interpretation of this finding is that in low complexity tasks, users may have set lower expectations regarding the response quality; as such, when the responses' quality was compromised, it did not greatly influence their

perception of quality and subsequently they overestimated it. An alternative explanation could be that participants who had less experience with chatbot technology could have overemphasized the responses' quality. A similar idea that might explain this finding is the positive correlation that we found between social orientation towards chatbot technology (Cf. section 4.4) and perceived quality. Thus, it is possible that participants who had participated in these conditions perceived chatbots more as social beings in lieu of machines, and consequently perceived the response quality to be higher. To confirm this, we performed a t-test to examine the difference in participants' self-reported social orientation scores between low and high-complexity tasks. We found that when the task complexity was low, participants' self-reported social orientation scores were greater (3.1 ± 1.65) as compared to high complexity tasks (2.6 ± 1.68) ($t(235) = 2.276, p = .024$). Thus, we can conclude that participants who preferred more humanized social interaction with the CPCS, perceived the quality to be higher even when it was compromised.

5.4 RQ4: Task Types Matter: Users in the Stress Task were More Affected with Waiting

Participants who interacted with an IR-bot perceived delays to be shorter (cf. Figure 6.A) and perceived the quality of interaction to be more positive than those who interacted with the Stress-bot (cf. Figure 10.B). Why participants, in the IR task, had a better waiting experience could be associated with the nature of IR tasks where the only goal is to extract relevant information from the agent as opposed to engaging in multi-turn, unconstrained and continued dialogue with the agent to alleviate stress and maintain rapport. Thus, it is possible that participants in the IR task were able to tolerate some delays due to its short-lived, single-turn nature. Recent research has tended to show a similar trend where Peng et al. [89] studied the impact of different response delays (1, 2, 4, 8s) on the users' satisfaction, in a human-robot context, for two task types: chitchat and IR. Their results show that when the delays were longer (4s and 8s), users in the chitchat condition showed less satisfaction than those who engaged with the IR tasks.

We could relate our participants' positive perception of interaction quality with the IR task to the way we designed the contents of responses for both tasks. For example, in the stress task, answers were derived from templated responses that were rather generic in nature. Despite the fact that these answers were verified by professional psychologists, it is probable that some participants were dissatisfied with the responses due to their generic nature and lack of detailed and contextualized feedback. In the case of IR tasks, it was comparatively easier to design a high-quality response because the search tasks were pre-defined and were presented in the form of quick replies. Secondly, aggregating answers from multiple information sources to tell a coherent story was a straightforward task.

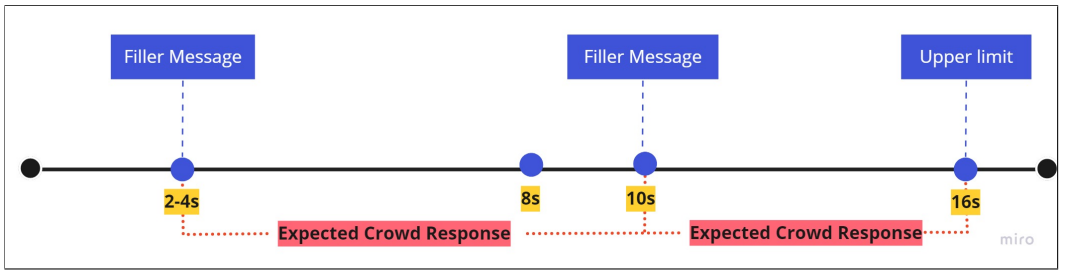

Fig. 12. Timeline: it indicates locations where the creator of CPCS can insert fillers to alleviate waiting feelings and indicates locations where a message from the crowd is expected.

5.5 Participants' Perceptions of the Bots' Identities

Given that we did not inquire participants about their perception regarding the nature of the chatbot in this study, this raises the question of what "identity" participants perceived toward the bots when they interacted with them - Did they view them as human or AI? In previous trials involving around 1000 participants, we discovered that approximately 73% of subjects believed they interacted with a form of hybrid intelligence, whereas the other 27% believed they interacted exclusively with AI [7]. Based on this, we expect that the current experiment would yield comparable results. It is noteworthy that even if participants believed that they were interacting with AI, this is to be expected in a hybrid CPCS setting where there is a role of AI in either creating a response or evaluating the quality of human-generated output. While current research on AI and conversational agents has focused on revealing the true nature of systems in laboratory trials due to ethical issues [99], masking the identity of a bot has shown some advantages. For instance, hybrid CPCSs are comparable to chatbots often employed in customer service contexts, where concealing the true nature of a conversational agent may be helpful in certain circumstances, particularly when seamless overlapping of human (customer representative) and AI is involved [110]. Another advantage of opacity was discovered in another study, where the bot was deployed to answer users' questions on the Stack Overflow community. They found that when the bot pretended to be a person (similar to our case), it was perceived more positively than when the bot revealed its bot identity [84]. Furthermore, both bots exhibit some features that may help to perceive them as human-like. For instance, a dynamic chatbot that generated relevant follow-up questions was seen as more human-like than a static bot that did not generate such inquiries [105]. This was true in the case of the Stress-bot, which uses MI theory to ask follow-up questions through multiple turns. However, in the case of an IR-bot, we used simple, polite, and informal language to create responses, which is indicative of humanness [8]. Our future research will examine how users' perceptions change when they learn post-hoc whether the bot was controlled by humans or AI.

6 IMPLICATIONS FOR DESIGNING FUTURE CROWD-POWERED CONVERSATIONAL SYSTEMS

We found that some participants in general did not like fast responses for two reasons: 1) they thought that CPCS had not sincerely pondered and understood their problem and was only concerned about answering quickly; 2) They believed that the answer was auto generated because it is impossible for a human to read and respond to a message in such a short time. Therefore, it seemed like postponing the answer up to 4s is not strenuous in the context of CPCS. In case of longer responses, one prominent solution that was proposed by participants was using filler or holding messages (e.g., "please bear with me while I think about that")–Figure 11. Thus, the designers of CPCS can add these filler messages after certain time period to let the users believe that CPCS is still actively listening, in addition to more traditional graphical indicators (Figure 12). We propose that after a 4s period, a filler message should be displayed to acknowledge the question, and to request the user to wait until receiving an answer from the crowd. Ideally, in the next 4 to 8s, an answer should be received from the crowd. If no answer is received within 8s, then the CPCS can wait for a couple of more seconds (based on the theory proposed by Nielsen [85]) for the answer. If no answer is received within 10s, another filler message can be displayed that can explain the reasons for the delay and ask the user to further wait.

Workers in CPCS can at times take longer than 16s to respond due to variety of reasons that are outside their control, such as slow typing speed, fatigue, a poor Internet connection and other technical issues. In these scenarios, filler messages can be augmented with other options, such as those proposed by participants in our study. For instance, in case of search tasks that take longer to

respond, CPCS can offer users to solve a quiz or a puzzle to sooth the effects of waiting. In case of stress tasks, it is more relevant to augment filler messages with calming music or relaxing visual images/animations that showed effectiveness in relieving stress [118]. Furthermore, since the CPCS are relying on humans, one can infer the expected waiting time based of the worker's typing speed and the average number of words that the worker has earlier written. Based on this, one can display a timer or progress bar to give a waiting estimate.

To ensure and maintain high quality input from crowd workers, control mechanisms have already been proposed [6]. For instance, Chorus [69] implements an explicit voting method where workers vote on the messages proposed by their fellow workers, while Evorus [48] implements an automatic voting technique based on machine learning models to improve the quality of responses over time. Although these methods help to improve the response quality to some extent, they can be ineffective. For example, in the case there are many spammers or inexperienced workers because the majority voting would weigh the votes of all workers equally [56]. For instance, it is a challenging task for an unskilled worker to provide effective coaching to people who are distressed. Such coaching requires workers to practice a plethora of skills ranging from understanding a person's thoughts and feelings to deciding what psychological interventions to provide based on the situation. In a recent study [3], researchers explored the efficacy of a conversational user interface (CUI) for training crowd workers to deliver complex therapeutic tasks based on MI skills. Results show that workers who were trained through CUI provided better psychological interventions and felt higher self-efficacy for dealing with stress management tasks. Similar CUIs can be employed to train workers to retrieve high-quality answers in case of complex information finding tasks. Additionally, one can nudge workers to enhance their learning gain for complex conversational tasks by setting learning goals for those who have high learning goal orientation [97]. These solutions can help to improve the content quality of responses, and can consequently improve the waiting experience and engagement.

Predicting complexity for complex conversational tasks could be beneficial for both workers and users of the CPCS system. For instance, if the CPCS is able to measure complexity of the utterance posed by the user, it can route the request to the workers who are fastest typists or assign the task to those who have the required skills or training to accomplish the task more quickly. Although various computation-based techniques have been developed to measure cognitive load of operators [60] or measuring the complexity of the individual messages in chatbots [39], it remains a challenge to assess complexity for diverse and complex conversational tasks. For example, in a stress task, merely measuring complexity based on the language used would not be sufficient since the task may involve a variety of stressors that would then define the complexity. Therefore, there is a need to quantify the stressful state context in stress mitigation tasks [88]. Similarly, IR tasks may also involve various dimensions that can define their complexity, such as the number of information resources needed to forge, the number of activities involved, the number of steps required and the expected workers' search behaviors in term of the query length, the number of URLs visited, etc [57]. Yang et al. [124] employed an extensive approach to quantify subjective complexity of crowdsourcing tasks based on some measurable properties of tasks, such as metadata features (e.g., title, descriptions etc.), content features (e.g., words count, links, and images etc.), and visual features (e.g., colorfulness, stylesheet etc.). Their proposed method was able to use these features to accurately predict the task complexity and task performance. Similar approaches are needed to quantify complexity of crowdsourced conversational tasks.

In summary, our study highlights several key findings and design implications for CPCSs that are presented below:

- ★ Users increasingly underestimated the waiting time when the delay was above 2s and were able to tolerate the delay up to 8s. Thus, researchers need to develop algorithms and workflows to keep the response latency within 4-8 seconds for CPCSs.
- ★ CPCSs can be equipped with different conversational fillers or acknowledgement tokens to mitigate the waiting time in CPCSs. For exceedingly longer delays, some non-temporal tasks such as solving a puzzle or an animation can be used to distract users.
- ★ Since our perception is our reality, we can aim for lowering the subjective waiting time in CPCS by uplifting the quality of response contents through workers' training, which does not require much engineering work as opposed to algorithmically lowering objective waiting time or system response time in CPCS.
- ★ An inverse relationship between task complexity and time perception holds within CPCS. Thus, computationally predicting the complexity for crowdsourced conversational tasks could be helpful in less complicated tasks to route the request to workers who are fastest typists or assign the task to those who have the required skills to accomplish the task more quickly.
- ★ People are less affected with waiting in task-centric conversations that are shallow in focus and need fewer exchanges compared to emotion-centric conversations that are deep in focus and require multiple exchanges in CPCS. In the former case, strict time constraints can be eased for crowd workers to generate answers.

7 LIMITATIONS

A limitation of the study concerns the nature of user responses. These were restricted to choosing between template-based utterances rather than typing in free responses. While this approach is adopted for reasons of ease of use in many chatbot applications, it is also not universally applicable especially in the cases where a free form conversation is intended, as is the case in cases of dealing with stress or other personal problems. Given the restricted nature of text entry in our experiments, our results may not easily generalize to cases of extended free form text input, where perhaps a largely varying time to compose an input or even the content of the system response may influence the expectations of the user regarding the response delay. Future studies could also address if the type of time-filler selected for different cases is considered suitable in different context and for different types of conversation. However, our study does inform what type of "filler" would be applicable for a certain latency.

We purposefully introduced complexity into the design to test the hypothesis that when complexity of the user input is low, participants will have less patience for extended delays than when complexity is high. As a result, we did not adjust the bot's response time based on the complexity of the user enquiry to determine how this would affect the perception of time. However, given the stress task's free-form nature, it is possible that users will occasionally enter simple responses (such as "hello") anticipating a prompt response. We controlled this concern in two ways in the Stress-bot: 1) we designed it in accordance with motivational interviewing literature, which encourages participants to ponder and enter practical responses rather than short inquiries (e.g., what concerns you about your problem?); 2) The Stress-bot initiated the conversation and greeted users, avoiding short inquiries from users, and asked a question in the beginning to break the ice (e.g., Tell me what brought you here today?). In the case of an IR-bot, the length of the user's response had no bearing on the outcome, as the user's response had been pre-programmed. We intend to do

additional research to determine how varied lengths of user requests affect people's perceptions of CPCS delays in real-world contexts.

Another limitation of our study is that we cannot ascertain how participants interpreted bots' identities when they interacted with them. For instance, were they aware that the bots were pre-programmed with pre-defined responses or that the responses were augmented in real time by human intelligence? Future research will look into how the results vary between those who felt they were interacting with AI alone versus those who thought they were interacting with a bot powered by human intelligence.

ACKNOWLEDGMENTS

This work was funded partially by Mirpur University of Science and Technology, Mirpur AJK, Pakistan, under award number 611-19 P.D/2017, and in part by the Design@Scale Delft AI Lab, the 4TU.CEE UNCAGE project, and Eindhoven University of Technology, Netherlands. We would like to express our gratitude to the reviewers for their helpful feedback, as well as to all participants in our study.

REFERENCES

[1] Tahir Abbas, Ujwal Gadiraju, Vassilis-Javed Khan, and Panos Markopoulos. 2021. Making Time Fly: Using Fillers to Improve Perceived Latency in Crowd-Powered Conversational Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9, 1 (Oct. 2021), 2–14. https://ojs.aaai.org/index.php/HCOMP/article/view/18935

[2] Tahir Abbas, Vassilis-Javed Khan, Ujwal Gadiraju, Emilia Barakova, and Panos Markopoulos. 2020. Crowd of Oz: A Crowd-Powered Social Robotics System for Stress Management. *Sensors* 20, 2 (2020). https://doi.org/10.3390/s20020569

[3] Tahir Abbas, Vassilis-Javed Khan, Ujwal Gadiraju, and Panos Markopoulos. 2020. Trainbot: A Conversational Interface to Train Crowd Workers for Delivering On-Demand Therapy. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8, 1 (Oct. 2020), 3–12. https://ojs.aaai.org/index.php/HCOMP/article/view/7458

[4] Ritu Agarwal and Elena Karahanna. 2000. Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS quarterly* (2000), 665–694.

[5] Hee-Kyung Ahn, Maggie Wenjing Liu, and Dilip Soman. 2006. Memory for Time: A Cognitive Model of Retrospective Duration and Sequence Judgments. *Available at SSRN 897933* (2006).

[6] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar. 2013. Quality Control in Crowdsourcing Systems: Issues and Directions. *IEEE Internet Computing* 17, 2 (2013), 76–81. https://doi.org/10.1109/MIC.2013.20

[7] Anonymous. 2021. Making Time Fly: Using Fillers to Improve Perceived Latency in Crowd-Powered Conversational Systems. (2021). https://osf.io/b4ket/?view_only=8e6b6ee82c764a678dba237dce265e64

[8] Theo Araujo. 2018. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior* 85 (2018), 183–189. https://doi.org/10.1016/j.chb.2018.03.051

[9] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300484

[10] Ja Baptista-Neto, Fx Gingele, and T Leipe. [n.d.]. ZAR, JH, 1996. Biostatistical analysis, 3rd ed., New Jersey. *Environmental Pollution* 109 ([n. d.]), 1–9.

[11] Michael S. Bernstein, Joel Brandt, Robert C. Miller, and David R. Karger. 2011. Crowds in Two Seconds: Enabling Realtime Crowd-Powered Interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) *(UIST '11)*. Association for Computing Machinery, New York, NY, USA, 33–42. https://doi.org/10.1145/2047196.2047201

[12] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. 2010. VizWiz: Nearly Real-Time Answers to Visual Questions. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology* (New York, New York, USA) *(UIST '10)*. Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/1866029.1866080

[13] Florian Block and Hans Gellersen. 2010. The Impact of Cognitive Load on the Perception of Time. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (Reykjavik, Iceland) *(NordiCHI '10)*.

Association for Computing Machinery, New York, NY, USA, 607–610. https://doi.org/10.1145/1868914.1868985

[14] Richard A Block, Peter A Hancock, and Dan Zakay. 2000. Sex differences in duration judgments: A meta-analytic review. *Memory & Cognition* 28, 8 (2000), 1333–1346.

[15] Margaret Boden, Joanna Bryson, Darwin Caldwell, Kerstin Dautenhahn, Lilian Edwards, Sarah Kember, Paul Newman, Vivienne Parry, Geoff Pegman, Tom Rodden, Tom Sorrell, Mick Wallis, Blay Whitby, and Alan Winfield. 2017. Principles of robotics: regulating robots in the real world. *Connection Science* 29, 2 (2017), 124–129. https://doi.org/10.1080/09540091.2016.1271400 arXiv:https://doi.org/10.1080/09540091.2016.1271400

[16] Ria Mae Borromeo, Thomas Laurent, and Motomichi Toyama. 2016. The Influence of Crowd Type and Task Complexity on Crowdsourced Work Quality. In *Proceedings of the 20th International Database Engineering &; Applications Symposium* (Montreal, QC, Canada) *(IDEAS '16)*. Association for Computing Machinery, New York, NY, USA, 70–76. https://doi.org/10.1145/2938503.2938511

[17] Virginia Braun and Victoria Clarke. 2012. Thematic Analysis. *Handbook of Research Methods in psychology* 2 (2012). https://doi.org/10.1037/13620-004

[18] Thomas W Butler. 1983. Computer response time and user performance.. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 58–62.

[19] Jessy Ceha, Ken Jen Lee, Elizabeth Nilsen, Joslin Goh, and Edith Law. 2021. *Can a Humorous Conversational Agent Enhance Learning Experience and Outcomes?* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445068

[20] Ki Hun Cho, Min Kyu Kim, Hwang-Jae Lee, and Wan Hee Lee. 2015. Virtual reality training with cognitive load improves walking function in chronic stroke patients. *The Tohoku journal of experimental medicine* 236, 4 (2015), 273–280.

[21] Alan J Christensen, Patricia J Moran, John S Wiebe, Shawna L Ehlers, and William J Lawton. 2002. Effect of a behavioral self-regulation intervention on patient adherence in hemodialysis. *Health Psychology* 21, 4 (2002), 393.

[22] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.

[23] Benilda Eleonor V Comendador, Bien Michael B Francisco, Jefferson S Medenilla, and Sharleen Mae. 2015. Pharmabot: a pediatric generic medicine consultant chatbot. *Journal of Automation and Control Engineering* 3, 2 (2015).

[24] Mihaly Czikszentmihalyi. 1990. *Flow: The psychology of optimal experience*. New York: Harper & Row.

[25] Tom Van Daele, Dirk Hermans, Chantal Van Audenhove, and Omer Van den Bergh. 2012. Stress Reduction Through Psychoeducation: A Meta- Analytic Review. *Health Education & Behavior* 39, 4 (2012), 474–485. https://doi.org/10.1177/1090198111419202 arXiv:https://doi.org/10.1177/1090198111419202 PMID: 21986242.

[26] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Comput. Surv.* 51, 1, Article 7 (Jan. 2018), 40 pages. https://doi.org/10.1145/3148148

[27] Mark M Davis and Janelle Heineke. 1998. How disconfirmation, perception and actual waiting times impact customer satisfaction. *international Journal of Service industry Management* (1998).

[28] Zoltan Dienes. 2014. Using Bayes to get the most out of non-significant results. *Frontiers in psychology* 5 (2014), 781.

[29] Kent Drummond and Robert Hopper. 1993. Acknowledgment tokens in series. *Communication Reports* 6, 1 (1993), 47–53. https://doi.org/10.1080/08934219309367561 arXiv:https://doi.org/10.1080/08934219309367561

[30] Yingling Fan, Andrew Guthrie, and David Levinson. 2016. Waiting time perceptions at transit stops and stations: Effects of basic amenities, gender, and security. *Transportation Research Part A: Policy and Practice* 88 (2016), 251–264. https://doi.org/10.1016/j.tra.2016.04.012

[31] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[32] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.

[33] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467. https://doi.org/10.1080/10447318.2018.1456150 arXiv:https://doi.org/10.1080/10447318.2018.1456150

[34] Norman M. Fraser and G.Nigel Gilbert. 1991. Simulating speech systems. *Computer Speech & Language* 5, 1 (1991), 81–99. https://doi.org/10.1016/0885-2308(91)90019-M

[35] Peter Fröhlich. 2005. Dealing with system response times in interactive speech applications. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*. 1379–1382.

[36] Ujwal Gadiraju and Stefan Dietze. 2017. Improving Learning through Achievement Priming in Crowdsourced Information Finding Microtasks. In *Proceedings of the Seventh International Learning Analytics &; Knowledge Conference* (Vancouver, British Columbia, Canada) *(LAK '17)*. Association for Computing Machinery, New York, NY, USA, 105–114. https://doi.org/10.1145/3027385.3027402

[37] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2017. Clarity is a Worthwhile Quality: On the Role of Task Clarity in Microtask Crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media* (Prague, Czech Republic) *(HT '17)*. Association for Computing Machinery, New York, NY, USA, 5–14. https://doi.org/10.1145/3078714.3078715

[38] Joseph Glicksohn. 2001. Temporal Cognition and the Phenomenology of Time: A Multiplicative Function for Apparent Duration. *Consciousness and Cognition* 10, 1 (2001), 1–25. https://doi.org/10.1006/ccog.2000.0468

[39] Ulrich Gnewuch, Stefan Morana, Marc Adam, and Alexander Maedche. 2018. Faster is not always better: understanding the effect of dynamic response delays in human-chatbot interaction. (2018).

[40] Gerald J Gorn, Amitava Chattopadhyay, Jaideep Sengupta, and Shashank Tripathi. 2004. Waiting for the web: how screen color affects time perception. *Journal of marketing research* 41, 2 (2004), 215–225.

[41] Karen Grace-Martin. 2020. When Unequal Sample Sizes Are and Are NOT a Problem in ANOVA - The Analysis Factor. https://www.theanalysisfactor.com/when-unequal-sample-sizes-are-and-are-not-a-problem-in-anova/

[42] Jonathan Grudin and Richard Jacques. 2019. Chatbots, Humbots, and the Quest for Artificial General Intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300439

[43] Daniel Haas, Jiannan Wang, Eugene Wu, and Michael J. Franklin. 2015. CLAMShell: Speeding up Crowds for Low-Latency Data Labeling. *Proc. VLDB Endow.* 9, 4 (Dec. 2015), 372–383. https://doi.org/10.14778/2856318.2856331

[44] John A Hoxmeier and Chris DiCesare. 2000. System response time and user satisfaction: An experimental study of browser-based applications. *AMCIS 2000 Proceedings* (2000), 347.

[45] Ting-Hao Huang and Jeffrey Bigham. 2017. A 10-Month-Long Deployment Study of On-Demand Recruiting for Low-Latency Crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 5, 1 (Sep. 2017). https://ojs.aaai.org/index.php/HCOMP/article/view/13318

[46] Ting-Hao Huang, Walter Lasecki, Amos Azaria, and Jeffrey Bigham. 2016. "Is There Anything Else I Can Help You With?": Challenges in Deploying an On-Demand Crowd-Powered Conversational Agent. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 4, 1 (Sep. 2016), 79–88. https://ojs.aaai.org/index.php/HCOMP/article/view/13292

[47] Ting-Hao Huang, Walter Lasecki, and Jeffrey Bigham. 2015. Guardian: A Crowd-Powered Spoken Dialog System for Web APIs. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 3, 1 (Sep. 2015). https://ojs.aaai.org/index.php/HCOMP/article/view/13237

[48] 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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173869

[49] Ting-Hao K. Huang, Amos Azaria, Oscar J. Romero, and Jeffrey P. Bigham. 2019. InstructableCrowd: Creating IF-THEN Rules for Smartphones via Conversations with the Crowd. *Hum. Comput.* 6 (2019), 113–146. https://doi.org/10.15346/hc.v6i1.7

[50] Ting-Hao K. Huang, Yun-Nung Chen, and Jeffrey P. Bigham. 2017. Real-time On-Demand Crowd-powered Entity Extraction. *CoRR* abs/1704.03627 (2017). arXiv:1704.03627 http://arxiv.org/abs/1704.03627

[51] Michael K. Hui and David K. Tse. 1996. What to Tell Consumers in Waits of Different Lengths: An Integrative Model of Service Evaluation. *Journal of Marketing* 60, 2 (1996), 81–90. https://doi.org/10.1177/002224299606000206 arXiv:https://doi.org/10.1177/002224299606000206

[52] Panagiotis G. Ipeirotis. 2010. Analyzing the Amazon Mechanical Turk Marketplace. *XRDS* 17, 2 (Dec. 2010), 16–21. https://doi.org/10.1145/1869086.1869094

[53] Julie A Jacko, Andrew Sears, and Michael S Borella. 2000. The effect of network delay and media on user perceptions of web resources. *Behaviour & Information Technology* 19, 6 (2000), 427–439.

[54] Gunn Johansson and Gunnar Aronsson. 1984. Stress reactions in computerized administrative work. *Journal of organizational behavior* 5, 3 (1984), 159–181.

[55] Sepandar D Kamvar and Jonathan Harris. 2011. We feel fine and searching the emotional web. In *Proceedings of the fourth ACM international conference on Web search and data mining*. 117–126.

[56] David R Karger, Sewoong Oh, and Devavrat Shah. 2011. Iterative learning for reliable crowdsourcing systems. Neural Information Processing Systems, 1953–1961.

[57] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. 2015. Development and Evaluation of Search Tasks for IIR Experiments Using a Cognitive Complexity Framework. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval* (Northampton, Massachusetts, USA) *(ICTIR '15)*. Association for Computing Machinery, New York, NY, USA, 101–110. https://doi.org/10.1145/2808194.2809465

[58] Deanna S. Kempf. 1999. Attitude formation from product trial: Distinct roles of cognition and affect for hedonic and functional products. *Psychology & Marketing* 16, 1 (1999), 35–50. https://doi.org/10.1002/(SICI)1520-6793(199901)16:

1<35::AID-MAR3>3.0.CO;2-U

[59] Azizuddin Khan, Narendra K Sharma, and Shikha Dixit. 2006. Effect of cognitive load and paradigm on time perception. *Journal of the Indian Academy of applied Psychology* 32, 1 (2006), 37–42.

[60] M. Asif Khawaja, Fang Chen, and Nadine Marcus. 2010. Using Language Complexity to Measure Cognitive Load for Adaptive Interaction Design. In *Proceedings of the 15th International Conference on Intelligent User Interfaces* (Hong Kong, China) *(IUI '10)*. Association for Computing Machinery, New York, NY, USA, 333–336. https://doi.org/10.1145/1719970.1720024

[61] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 453–456. https://doi.org/10.1145/1357054.1357127

[62] Kira Kretzschmar, Holly Tyroll, Gabriela Pavarini, Arianna Manzini, Ilina Singh, and NeurOx Young People's Advisory Group. 2019. Can Your Phone Be Your Therapist? Young People's Ethical Perspectives on the Use of Fully Automated Conversational Agents (Chatbots) in Mental Health Support. *Biomedical Informatics Insights* 11 (2019), 1178222619829083. https://doi.org/10.1177/1178222619829083 arXiv:https://doi.org/10.1177/1178222619829083 PMID: 30858710.

[63] Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology* 4 (2013), 863.

[64] Carine Lallemand and Guillaume Gronier. 2012. Enhancing User EXperience during Waiting Time in HCI: Contributions of Cognitive Psychology. In *Proceedings of the Designing Interactive Systems Conference* (Newcastle Upon Tyne, United Kingdom) *(DIS '12)*. Association for Computing Machinery, New York, NY, USA, 751–760. https://doi.org/10.1145/2317956.2318069

[65] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-Time Captioning by Groups of Non-Experts. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (Cambridge, Massachusetts, USA) *(UIST '12)*. Association for Computing Machinery, New York, NY, USA, 23–34. https://doi.org/10.1145/2380116.2380122

[66] Walter S Lasecki, Christopher Homan, and Jeffrey P Bigham. 2014. Architecting real-time crowd-powered systems. *Human Computation* 1, 1 (2014).

[67] Walter S. Lasecki, Kyle I. Murray, Samuel White, Robert C. Miller, and Jeffrey P. Bigham. 2011. Real-Time Crowd Control of Existing Interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) *(UIST '11)*. Association for Computing Machinery, New York, NY, USA, 23–32. https://doi.org/10.1145/2047196.2047200

[68] Walter S. Lasecki, Phyo Thiha, Yu Zhong, Erin Brady, and Jeffrey P. Bigham. 2013. Answering Visual Questions with Conversational Crowd Assistants. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility* (Bellevue, Washington) *(ASSETS '13)*. Association for Computing Machinery, New York, NY, USA, Article 18, 8 pages. https://doi.org/10.1145/2513383.2517033

[69] Walter S. Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F. Allen, and Jeffrey P. Bigham. 2013. Chorus: A Crowd-Powered Conversational Assistant. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (St. Andrews, Scotland, United Kingdom) *(UIST '13)*. Association for Computing Machinery, New York, NY, USA, 151–162. https://doi.org/10.1145/2501988.2502057

[70] Younghwa Lee, Andrew N. K. Chen, and Virginia Ilie. 2012. Can Online Wait Be Managed? The Effect of Filler Interfaces and Presentation Modes on Perceived Waiting Time Online. *MIS Quarterly* 36, 2 (2012), 365–394. http://www.jstor.org/stable/41703460

[71] Q. Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N. Sadat Shami. 2016. What Can You Do? Studying Social-Agent Orientation and Agent Proactive Interactions with an Agent for Employees. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (Brisbane, QLD, Australia) *(DIS '16)*. Association for Computing Machinery, New York, NY, USA, 264–275. https://doi.org/10.1145/2901790.2901842

[72] Jake Linardon and Matthew Fuller-Tyszkiewicz. 2020. Attrition and adherence in smartphone-delivered interventions for mental health problems: A systematic and meta-analytic review. *Journal of consulting and clinical psychology* 88, 1 (2020), 1.

[73] Andrey Lovakov and Elena Agadullina. 2017. Empirically derived guidelines for interpreting effect size in social psychology. (2017).

[74] Kate MacFarlane and Lisa Marks Woolfson. 2013. Teacher attitudes and behavior toward the inclusion of children with social, emotional and behavioral difficulties in mainstream schools: An application of the theory of planned behavior. *Teaching and teacher education* 29 (2013), 46–52.

[75] Robert B Miller. 1968. Response time in man-computer conversational transactions. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*. 267–277.

[76] Robert B. Miller. 1968. Response Time in Man-Computer Conversational Transactions. In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I* (San Francisco, California) *(AFIPS '68 (Fall, part I))*. Association for Computing Machinery, New York, NY, USA, 267–277. https://doi.org/10.1145/1476589.1476628

[77] William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. Manual for the motivational interviewing skill code (MISC). *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico* (2003).

[78] William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.

[79] Kathy Missildine, Rebecca Fountain, Lynn Summers, and Kevin Gosselin. 2013. Flipping the classroom to improve student performance and satisfaction. *Journal of Nursing Education* 52, 10 (2013), 597–599.

[80] Youngme Moon. 1999. The effects of physical distance and response latency on persuasion in computer-mediated communication and human–computer communication. *Journal of Experimental Psychology: Applied* 5, 4 (1999), 379.

[81] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100.

[82] Robert Morris. 2011. Crowdsourcing workshop: the emergence of affective crowdsourcing. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems. ACM*. Citeseer.

[83] Robert Robert Randall Morris. 2015. *Crowdsourcing mental health and emotional well-being*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[84] Alessandro Murgia, Daan Janssens, Serge Demeyer, and Bogdan Vasilescu. 2016. Among the Machines: Human-Bot Interaction on Social Q&A Websites. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 1272–1279. https://doi.org/10.1145/2851581.2892311

[85] Jakob Nielsen. 1994. *Usability engineering*. Morgan Kaufmann.

[86] University of Gothenburg. 2015. Pauses can make or break a conversation. www.sciencedaily.com/releases/2015/09/150930110555.htm. Accessed: 2021-03-18.

[87] Adam Palanica, Peter Flaschner, Anirudh Thommandram, Michael Li, and Yan Fossat. 2019. Physicians' Perceptions of Chatbots in Health Care: Cross-Sectional Web-Based Survey. *J Med Internet Res 2019;21(4):e12887 https://www.jmir.org/2019/4/e12887* 21, 4 (apr 2019), e12887. https://doi.org/10.2196/12887

[88] SoHyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee, and Bongwon Suh. 2019. Designing a Chatbot for a Brief Motivational Interview on Stress Management: Qualitative Case Study. *J Med Internet Res* 21, 4 (16 Apr 2019), e12231. https://doi.org/10.2196/12231

[89] Zhenhui Peng, Kaixiang Mo, Xiaogang Zhu, Junlin Chen, Zhijun Chen, Qian Xu, and Xiaojuan Ma. 2020. Understanding User Perceptions of Robot's Delay, Voice Quality-Speed Trade-off and GUI during Conversation. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3382792

[90] Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What Makes a Good Counselor? Learning to Distinguish between High-quality and Low-quality Counseling Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 926–935. https://doi.org/10.18653/v1/P19-1088

[91] Olga Perski, Ann Blandford, Robert West, and Susan Michie. 2017. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. *Translational behavioral medicine* 7, 2 (2017), 254–267.

[92] Olga Perski, David Crane, Emma Beard, and Jamie Brown. 2019. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. *Digital health* 5 (2019), 2055207619880676.

[93] Ad Pruyn and Ale Smidts. 1998. Effects of waiting on the satisfaction with the service: Beyond objective time measures1Both authors contributed equally to this article.1. *International Journal of Research in Marketing* 15, 4 (1998), 321 – 334. https://doi.org/10.1016/S0167-8116(98)00008-1

[94] A Th H Pruyn and Ale Smidts. 1993. Customers' evaluations of queues: Three exploratory studies. *ACR European Advances* (1993).

[95] 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376403

[96] 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*. 1–5.

[97] Amy Rechkemmer and Ming Yin. 2020. Motivating Novice Crowd Workers through Goal Setting: An Investigation into the Effects on Complex Crowdsourcing Task Training. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8, 1 (Oct. 2020), 122–131. https://ojs.aaai.org/index.php/HCOMP/article/view/7470

[98] René Riedl and Thomas Fischer. 2018. System response time as a stressor in a digital world: literature review and theoretical model. In *International Conference on HCI in Business, Government, and Organizations*. Springer, 175–186.

[99] Laurel D. Riek. 2012. Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *J. Hum.-Robot Interact.* 1, 1 (July 2012), 119–136. https://doi.org/10.5898/JHRI.1.1.Riek

[100] Carl R Rogers and Richard E Farson. 1957. *Active listening*. Chicago, IL.

[101] Emma J. Rose and Elin A. Björling. 2017. Designing for Engagement: Using Participatory Design to Develop a Social Robot to Measure Teen Stress. In *Proceedings of the 35th ACM International Conference on the Design of Communication* (Halifax, Nova Scotia, Canada) *(SIGDOC '17)*. Association for Computing Machinery, New York, NY, USA, Article 7, 10 pages. https://doi.org/10.1145/3121113.3121212

[102] Ayelet Meron Ruscio and Dana Rabois Holohan. 2006. Applying empirically supported treatments to complex cases: Ethical, empirical, and practical considerations. *Clinical Psychology: Science and Practice* 13, 2 (2006), 146–162.

[103] Denis Savenkov and Eugene Agichtein. 2016. CRQA: Crowd-Powered Real-Time Automatic Question Answering System. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 4, 1 (Sep. 2016). https://ojs.aaai.org/index.php/HCOMP/article/view/13291

[104] Lawrence M Schleifer and Benjamin C Amick III. 1989. System response time and method of pay: Stress effects in computer-based tasks. *International Journal of Human-Computer Interaction* 1, 1 (1989), 23–39.

[105] Ryan M Schuetzler, Mark Grimes, Justin Scott Giboney, and Joesph Buckman. 2014. Facilitating natural conversational agent interactions: lessons from a deception experiment. (2014).

[106] Stela H Seo, Keelin Griffin, James E Young, Andrea Bunt, Susan Prentice, and Verónica Loureiro-Rodríguez. 2018. Investigating people's rapport building and hindering behaviors when working with a collaborative robot. *International Journal of Social Robotics* 10, 1 (2018), 147–161.

[107] Steven C Seow. 2008. *Designing and engineering time: The psychology of time perception in software*. Addison-Wesley Professional.

[108] T. Shiwa, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita. 2008. How quickly should communication robots respond?. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 153–160. https://doi.org/10.1145/1349822.1349843

[109] Ben Shneiderman. 1984. Response time and display rate in human performance with computers. *ACM Computing Surveys (CSUR)* 16, 3 (1984), 265–285.

[110] Marita Skjuve, Ida Maria Haugstveit, Asbjørn Følstad, and Petter Bae Brandtzaeg. 2019. Help! Is my chatbot falling into the uncanny valley? An empirical study of user experience in human-chatbot interaction. *Human Technology* 15, 1 (2019).

[111] T Smeets, J Leppink, M Jelicic, and HLGJ Merckelbach. 2009. Shortened versions of the Gudjonsson Suggestibility Scale meet the standards. *Legal and criminological Psychology* 14, 1 (2009), 149–155.

[112] Shirley Taylor. 1994. The Effects of Filled Waiting Time and Service Provider Control over the Delay on Evaluations of Service. *Journal of the Academy of Marketing Science* 23, 1 (1994), 38–48. https://doi.org/10.1177/0092070395231005 arXiv:https://doi.org/10.1177/0092070395231005

[113] Ewart AC Thomas and Wanda B Weaver. 1975. Cognitive processing and time perception. *Perception & psychophysics* 17, 4 (1975), 363–367.

[114] Michael Toomim, Travis Kriplean, Claus Pörtner, and James Landay. 2011. Utility of human-computer interactions: Toward a science of preference measurement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2275–2284.

[115] Janet Treasure. 2004. Motivational interviewing. *Advances in Psychiatric Treatment* 10, 5 (2004), 331–337.

[116] Hans Van Der Heijden. 2002. On the cognitive-affective structure of attitudes toward information systems. *ICIS 2002 Proceedings* (2002).

[117] Andrea E Waylen, Mark S Horswill, Jane L Alexander, and Frank P McKenna. 2004. Do expert drivers have a reduced illusion of superiority? *Transportation research part F: traffic psychology and behaviour* 7, 4-5 (2004), 323–331.

[118] Taelyr Weekly, Nicole Walker, Jill Beck, Sean Akers, and Meaghann Weaver. 2018. A Review of Apps for Calming, Relaxation, and Mindfulness Interventions for Pediatric Palliative Care Patients. *Children* 5, 2 (2018). https://doi.org/10.3390/children5020016

[119] World Health Organization (WHO). 2017. Mental health: massive scale-up of resources needed if global targets are to be met. https://www.who.int/news/item/06-06-2018-mental-health-massive-scale-up-of-resources-needed-if-global-targets-are-to-be-met

[120] Alex C. Williams, Harmanpreet Kaur, Gloria Mark, Anne Loomis Thompson, Shamsi T. Iqbal, and Jaime Teevan. 2018. *Supporting Workplace Detachment and Reattachment with Conversational Intelligence*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi-org.tudelft.idm.oclc.org/10.1145/3173574.3173662

[121] Philipp Wintersberger, Tobias Klotz, and Andreas Riener. 2020. Tell Me More: Transparency and Time-Fillers to Optimize Chatbots' Waiting Time Experience. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society* (Tallinn, Estonia) *(NordiCHI '20)*. Association for Computing Machinery, New York, NY, USA, Article 76, 6 pages. https://doi.org/10.1145/3419249.3420170

[122] Robert E Wood. 1986. Task complexity: Definition of the construct. *Organizational behavior and human decision processes* 37, 1 (1986), 60–82.

[123] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A New Chatbot for Customer Service on Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3506–3510. https://doi.org/10.1145/3025453.3025496

[124] Jie Yang, Judith Redi, Gianluca Demartini, and Alessandro Bozzon. 2016. Modeling Task Complexity in Crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 4, 1 (Sep. 2016). https://ojs.aaai.org/index.php/HCOMP/article/view/13283

[125] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300772

[126] Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences* (2020), 1–17.

[127] Yu Zhong, Walter S. Lasecki, Erin Brady, and Jeffrey P. Bigham. 2015. RegionSpeak: Quick Comprehensive Spatial Descriptions of Complex Images for Blind Users. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2353–2362. https://doi.org/10.1145/2702123.2702437

A SAMPLE STRESS MANAGEMENT TASK

Carla is a 21-year-old unmarried girl. She is a final year bachelor student studying Innovation Sciences. She is a very hard-working student and has received an 'A' grade on all of her assignments. Her teachers believe that she is one of the best students in the class. A new student named Hanna, has recently moved from another class to Carla's class. Hanna is also a very talented student and received 'A' grades on all of her assignments. The presence of Hanna is stressing Carla out. Prior to Hanna's arrival, Carla considered herself to be the best student in class, with the best grades. Now she feels like she needs to compete with Hanna and is wary of losing out on the 'Best Student Award', an award presented by the University to the student with the most stellar academic record at the end of the study program. Previously, Carla found assignments to be easy and straightforward but now she is obsessed about her performance in the assignments and tends to worry about unnecessary details. For instance, she is constantly sending emails to her teachers to clarify the assignments and to make sure that she is doing well according to the instructions. She believes that if she fails to earn that award, then she would be a failure. She finds this situation increasingly overwhelming, and this makes her feel distressed and worried. When Carla is stressed out, she sits in a quiet place and tells herself that this award is something confined to her university, and nobody would know or care about it in the grand scheme of things. So, it would not really impact her future employment prospects or what her peers may think of her. This strategy is called self-talk, and this relieves her stress to some degree.