Using Worker Avatars to Improve Microtask Crowdsourcing

SIHANG QIU, Delft University of Technology, Netherlands
ALESSANDRO BOZZON, Delft University of Technology, Netherlands
MAX V. BIRK, Eindhoven University of Technology, Netherlands
UJWAL GADIRAJU, Delft University of Technology, Netherlands

The future of crowd work has been identified to depend on worker satisfaction, but we lack a thorough understanding of how worker satisfaction can be increased in microtask crowdsourcing. Prior work has shown that one solution is to build tasks that are engaging. To facilitate engagement, two methods that have received attention in recent HCI literature are the use of video games and conversational interfaces. While these are largely different techniques, they aim for the same goal of reducing worker burden and increasing engagement in a task. On one hand, video games have huge motivation potential and translating game design elements for motivational purposes has shown positive effects. Recent work in games research has shown that the use of player *avatars* is effective in fostering interest, enjoyment, and other aspects pertaining to intrinsic motivation. On the other hand, conversational interfaces have been argued to have advantages over traditional GUIs due to facilitating a more human-like interaction. "*Conversational*" microtasking has recently been proposed to improve worker engagement in microtask marketplaces. The contexts of games and crowd work are underlined by the need to motivate and engage participants, yet the potential of using worker avatars to promote self-identification and improve worker satisfaction in microtask crowdsourcing has remained unexplored. Addressing this knowledge gap, we carried out a between-subject study involving 360 crowd workers. We investigated how worker avatars influence quality related outcomes of workers and their perceived experience, in conventional web and novel conversational interfaces. We equipped workers with the functionality of customizing their avatars, and selecting characterizations for their avatars, to understand whether identifying with an avatar can increase the motivation of workers. We found that using worker avatars with conversational interfaces can effectively reduce cognitive workload and increase worker retention. Our results indicate the occurrence of similarity and wishful avatar identification in crowdsourcing. Our findings have important implications in alleviating workers' perceived workload and on the design of crowdsourcing microtasks.

CCS Concepts: • **Information systems → Crowdsourcing**; • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: Crowdsourcing, Chatbot, Avatar, Perceived workload, Motivation.

ACM Reference Format:
Sihang Qiu, Alessandro Bozzon, Max V. Birk, and Ujwal Gadiraju. 2021. Using Worker Avatars to Improve Microtask Crowdsourcing. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (May 2021), 28 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Authors' addresses: Sihang Qiu, Delft University of Technology, Netherlands, s.qiu-1@tudelft.nl; Alessandro Bozzon, Delft University of Technology, Netherlands, a.bozzon@tudelft.nl; Max V. Birk, Eindhoven University of Technology, Netherlands, m.v.birk@tue.nl; Ujwal Gadiraju, Delft University of Technology, Netherlands, u.k.gadiraju-1@tudelft.nl.

2573-0142/2021/5-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Microtasking marketplaces such as Amazon's Mechanical Turk[1] (AMT) and Prolific[2] are flourishing since crowdsourcing is widely being used to acquire human input at scale. Thousands of people around the globe rely on marketplaces such as AMT to earn their livelihood or as a secondary source of income [17]. We are also currently witnessing an unmistakable growth in AI, robotics and automation across various labour markets. The importance of human input to either build ground truth datasets or label training data in spurring on advances in AI has been well documented — for example, consider the role of ImageNet in catalysing advances in the field of Computer Vision [14]. Human input has been argued to be indispensable in this age of automation and the foreseeable future [26].

Highlighting the global potential of microtasking and the larger area of crowd work, the HCI community has identified that the future of crowd work depends on designing tasks and creating work that can achieve both organizational performance and worker satisfaction [40]. Although recent research has taken great strides in improving the organizational performance [13], relatively little work has focused on improving worker satisfaction across the microtask crowdsourcing landscape. The monotonous nature of human intelligence tasks (HITs) on microtasking platforms, coupled with sub-optimal task design, results in low worker satisfaction and engagement. This has been found to be a consequence of boredom or fatigue during task execution, and high drop-out and task abandonment rates [28, 54]. Designing tasks to engage and motivate workers is a potential solution [48, 68].

To increase participant engagement and satisfaction, the use of video games or employing conversational interfaces, are two methods that have received attention in recent HCI literature. Relevant work in the field of games research has shown that identifying with avatars can be effective in improving players' enjoyment and satisfaction [5, 66]. To improve worker engagement in microtask crowdsourcing, recent work has proposed the use of conversational interfaces for task execution [8, 51, 57]. The contexts of games and crowd work are underlined by the need to motivate and engage participants, yet the potential of using worker avatars to promote identification and improve worker satisfaction in microtask crowdsourcing has remained unexplored. This is important to investigate, since using worker avatars and assigning avatars characteristics or personality traits can increase identification [53, 66]. Avatar identification has been studied from three perspectives — similarity identification, embodied identification, and wishful identification. Similarity identification refers to the identification related to the similarity between the avatar and the user; embodied identification refers to the identification of the feeling whether (and to what extent) the user is inside the avatar; and wishful identification represents the identification of avatar characteristics that the user would like to have. Prior works have shown that avatar appearance and characteristics can affect similarity and wishful identification respectively [32, 33], whereas embodied identification demands more avatar operations and interactions, which is very common in video games but not essential in crowdsourcing.

To operationalize similarity identification and wishful identification, in this study we support workers in (a) building their own representations by customizing the appearance of their avatars, and (b) characterizing their avatars before they begin task execution, by selecting one out of three desirable worker characterizations drawn from related literature (*diligent worker, competent worker, balanced worker*) [21, 36]. Since the influence of worker avatars in crowd work has remained unexplored, we know little about their impact on both conventional task interfaces as well as novel

[1] https://www.mturk.com/
[2] https://www.prolific.co/

conversational interfaces. We thereby delve into this comparison through our work. In this paper, we address the following research questions:

RQ1: How do worker avatars affect worker experience and quality-related outcomes in conventional web and novel conversational interfaces?

RQ2: How can worker self-identification with their avatars be facilitated using avatar customization, and worker characterization of customized avatars?

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

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

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

Our findings have important implications in terms of reducing perceived workload and improving the sense of success through task design in microtask crowdsourcing. As argued by prior work, this can be crucial to the sustainability of microwork marketplaces [40].

2 RELATED WORK

We discuss related literature from four relevant standpoints – workers subjective experience, conversational agents, motivational game-design in crowdsourcing, and identification with avatars in games.

2.1 Subjective Perception in Crowdsourcing

The subjective perception of crowd workers plays important roles in microtask crowdsourcing with regard to many aspects such as the quality and the incentive. Previous works have studied subjective perceptions concerning worker engagement [19, 48, 57], moods [58, 68], satisfaction [51], and enjoyment [7, 10].

Worker engagement is crucial to microtask crowdsourcing since it has positive effects on building better relationships with crowd workers. Researchers have already noticed the importance of worker engagement and proposed methods to measure and predict it [48]. A previous work combined crowdsourcing with the process of learning [19], suggesting that both engagement and performance could be improved. Our recent work quantified worker engagement and worker retention, and showed that conversational interfaces could significantly better retain crowd workers [57]. The effort that workers make is also a major factor that can affect task execution time and cost. Cheng et al. proposed an effective way to measure worker efforts using "error-time area" (ETA), enabling a requester to rapidly evaluate the efficiency [9]. Recent studies have also shown that worker

moods could affect quality-related crowdsourced outcomes [58, 68]. Results revealed that workers in pleasant moods outperformed significantly over unpleasant workers.

Apart from engagement, effort, and mood, the worker performance could be affected by more complex factors. Kazai et al. studied the relationship between workers' personality traits and crowdsourcing outcomes [36]. Considering the properties of outcomes such as accuracy and speed, workers can be classified into five main categories — Spammer, Sloppy, Incompetent, Competent and Diligent. Prior works also investigated the feasibility of using self-assessments to measure rather complex subjective properties like logical reasoning competence [22] and cognitive skill [29] — these subjective properties can significantly affect crowdsourcing results. Using such self-assessments before task execution could be useful for performance prediction and task assignment.

However, we lack a thorough understanding of how workers' experience related to tasks and their mental workload can be improved. Therefore, we investigate the impact of our proposed methods on workers' perceived workload, as well as other common metrics including worker engagement and quality-related outcomes.

2.2 Conversational Microtask Crowdsourcing

Conversational agents and interfaces are useful in many aspects since they feature a more human-like interaction. Compared to traditional graphical user interfaces, recent studies have shown that users using conversational interfaces generally reported better experience in terms of trust, enjoyment, and engagement [2, 3, 16]. Conversational agents have been widely applied in a variety of domains such as customer service [25], healthcare [4, 44, 60], education [1, 63], and information search [46].

Researchers have also attempted to apply conversational interfaces in crowdsourcing, initially as a tool to answer questions about general knowledge [42]. The conversational interface named Chorus was responsible for receiving users' questions, publishing tasks on crowdsourcing platforms to collect answers for the questions, and finally sending aggregated answers back to users [42]. This has become a popular way of using conversational interfaces in crowdsourcing, inspiring several works [34, 41]. Curios Cat is a conversational agent (with an interface) designed to construct the knowledge base by acquiring data from crowds. The users of Curios Cat are supposed to provide data by chatting with the agent [8]. A previous study by Mavridis et al. used conversational agents in crowdsourcing marketplaces [51]. The experiment was conducted on the online crowdsourcing platform, where workers are re-directed to Telegram to complete the microtasks by chatting with the agent. Our recent work designed an HTML-based conversational interface for microtasking, which can be directly embedded on crowdsourcing platforms, saving the inconvenience of re-directing to other messaging applications [59]. Such Web-based conversational interfaces were shown to be effective for improving worker engagement.

Based on the findings of previous studies, our work also employs the conversational interface based on HTML for conversational microtask execution, and compares it with traditional Web interfaces, to investigate whether using worker avatars can improve crowd work across the interface types.

2.3 Gamification in Crowdsourcing

Gamification has been extensively used in the realm of crowdsourcing to make workers more motivated and engaged [55]. Following Flow theory [12], Eickhoff et al. designed a game to attract workers to execute Relevance Assessment tasks, resulting in lower cost and fewer malicious behaviors [20]. A prior work used competition-based design to improve worker performance in microtask crowdsourcing on the CrowdFlower platform [61]. Furthermore, using gamification to exploit worker motivation and interest to enable volunteering crowdsourcing has shown to

be feasible in prior studies [55]. A previous study developed an online collaborative game to effectively crowdsource protein structures [10]. Similar methods are also extensively used to inspire volunteers [7], increase enjoyment [43], or support activism [50].

Gamification has been shown to be effective in crowdsourcing. However, we learned that games from previous studies for crowdsourcing are all task-specific, meaning the game must be well designed to meet the requirements of simultaneously engaging workers and acquiring specific types of data. There is no common guidelines or tools for rapidly developing a game with little overhead. Inspired by gamification methods in crowdsourcing, we study avatar customization as a means to motivate workers. Avatar customization is a simple interface manipulation and has shown to increase task engagement [6].

2.4 Identification with Avatars

Avatars have been employed in many different areas, particularly in gaming systems. Prior work has showcased how and why players can be engaged in digital games [62], which is widely accepted by the researchers of relevant fields. The authors proposed self-determination theory (SDT) to explain the reason that games are usually engaging, and suggested that players would be intrinsically motivated if the game was designed to satisfy players' psychological needs of self-determination.

Based on the model of enjoyment [67], Trepte et al. studied competitiveness, player life satisfaction, and avatar identification in video games. They found a strong relationship between avatar identification and game enjoyment [66]. Apart from the effect of game enjoyment, prior work found that avatar customization itself could be engaging and valuable to players, after the authors carried out an interview study about the game World of Warcraft [47]. Furthermore, giving personality traits or even names could be important to increase identification while creating an avatar [11, 53, 66]. Fictional characters or avatars sometimes present what users or players wish to be. Hoffner et al. interviewed children and young adults about their favorite characters. Results indicated that both similarity identification (gender) and wishful identification (characteristic) existed in their favorite characters [32, 33]. Furthermore, Neustaedter et al. presented a study showing players created and evolved the avatar in games to match a desired virtual identity [56].

Based on the theories proposed and supported in previous works, we designed, implemented, and evaluated a novel function in microtask crowdsourcing – enabling workers to customize their avatar appearance, and selection of their desired avatar characteristics.

3 METHODS USING WORKER AVATARS FOR CROWDSOURCING

To answer **RQ1** and understand how effective worker avatars are, based on the type of interface, we designed worker avatars in both conventional web interfaces and novel conversational interfaces [57] for task execution. To answer **RQ2** and understand how avatar customization and characterization affects crowd work, we facilitated avatar customization and the selection of desired worker characterization across the web and conversational interfaces. To this end, we conducted a 3×2 between-subject study comparing three avatar conditions (without avatar, with avatar, with avatar and desirable characterizations) and two worker interfaces (Web and Chat), across two task types (Image Transcription and Information Finding). Addressing the **RQs**, we considered the following dependent variables – perceived workload, intrinsic motivation, and worker performance.

3.1 Avatar Design

3.1.1 Avatar Appearance. We used an avatar library called *avataaars*[3] to create 2D avatars by combining a variety of attributes, i.e., clothes, hair, emotions, accessories, and colors. Figure 1 (a)

[3] https://avataaars.com

shows the HTML-based panel for avatar appearance customization. In the avatar customization panel, we provided seven options for changing the avatar appearance: skin color, hair, facial hair, hair color, mood, accessories, and cloth color.

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

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

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

3) Mood. Eyes, eyebrow, and mouth types are initialized according to workers' moods. Since previous studies have shown the importance of worker moods in crowdsourcing [58, 68], we created a "mood" option by combining eyes, eyebrow, and mouth options. Note that in the original version of *avataaars*, the "mood" option does not exist – users need to customize moods by changing the emotion of eyes, eyebrow, and mouth. Using the mood option holistically instead of the individual attributes of eyes, eyebrow, and the mouth, we facilitate easy avatar appearance customization.

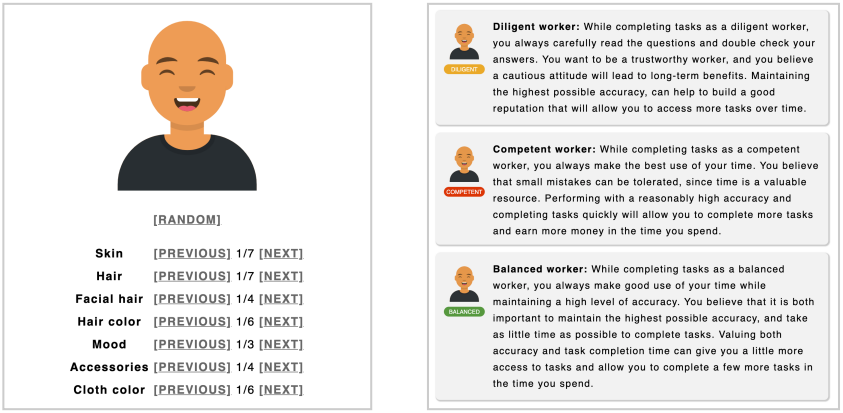
Apart from gender, skin, and mood, the accessories (types of glasses), and the color of the attire are randomly assigned for their initial avatar. After the avatar is initialized, workers have the freedom to change or randomize all previously mentioned options, as shown in Figure 1(a).

Considering that the most popular crowdsourcing marketplaces are Web-based, the avatars in our study are sketched on the Web-based interfaces using the vector format (SVG). Furthermore, the panel for avatar customization is purely based on HTML and JavaScript without any other dependencies. This makes the avatar customization very portable. Developers can easily deploy the avatar customization functionality to different Web applications with little overhead. The code repository for avatar appearance customization is shared publicly for the benefit of the community[4].

3.1.2 Avatar Characterizations. According to self-discrepancy theory [30], the "actual self" represents one's self-concept, while the "ideal self" is the representation of characteristics that one would like to have. By customizing the appearance of their avatars, workers can build their actual-self representations. Combined with ideal characteristics (for example, competence or diligence) workers can create a model avatar. The objective is to explore whether an avatar that workers self-identify with can also have characteristics that workers aspire to (wishful identification).

We provide three ideal characterizations for workers to select, as can be seen in Figure 1 (b). We adopted these characterizations from previous work [21, 36]. Authors synthesized the characteristics of online crowd workers and grouped them into five main categories — Diligent, Competent, Spammers, Less-competent, and Sloppy workers. In the original work by Kazai et al. [36], Diligent workers were characterized by a high ratio of high-quality output, longer average time spent per task, and high label accuracy. In comparison, Competent workers produce many useful labels and

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



(a) Appearance customization (b) Characterization selection

Fig. 1. The avatar (a) appearance customization and (b) characterization selection panels implemented based on HTML.

obtain high accuracies, but work relatively faster. Sloppy workers were characterized by their low task completion time and concomitant low accuracy. Incompetent workers are characterized by their high task completion times and concomitant low accuracy. Spammers were characterized by their ulterior motives to complete tasks quickly and maximize their rewards (by gaming the tasks), resulting in very low accuracies. Spammers, Less-competent, and Sloppy workers are negative characterizations that workers would want to avoid on crowdsourcing platforms — being perceived as sloppy might have negative consequences for workers, e.g., privileges are revoked or completed tasks are rejected without pay [54]. Due to the impact of rejection on worker reputation and their future access to tasks, workers typically refrain from wilfully under-performing in tasks. Therefore, as shown in Figure 1 (b) we do not consider negative characterizations and used the characterizations of **Diligent** and **Competent** workers in our study as those characteristics that workers aspire to (wishful identification). Considering the accuracy and completion time factors during task execution, Diligent workers are defined to exhibit high accuracy, but correspond to long task execution time, while Competent workers exhibit reasonably high accuracy and short task completion time [21]. In this study, we adopted the definitions of Diligent workers and Competent workers. But on the user interfaces shown to workers, we added more details about the motivation behind a worker characterization (i.e. why a worker can be diligent/competent). For example, we show "maintaining the highest possible accuracy can help to build a good reputation that will allow you to access more tasks over time" in the description of Diligent worker. We also introduced a **Balanced** characterization, to represent an ideal worker type who maintains balanced levels of accuracy and task execution speed that workers may wish to possess or aspire to.

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

- **Diligent worker.** *While completing tasks as a diligent worker, you always carefully read the questions and double-check your answers. You want to be a trustworthy worker, and you believe a cautious attitude will lead to long-term benefits. Maintaining the highest possible accuracy can help to build a good reputation that will allow you to access more tasks over time.*
- **Competent worker.** *While completing tasks as a competent worker, you always make the best use of your time. You believe that small mistakes can be tolerated, since time is a valuable resource.*

Performing with reasonably high accuracy and completing tasks quickly will allow you to complete more tasks and earn more money in the time you spend.

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

During the task execution, the selected and desired characterization is always displayed below the avatar in a `characterization label` (cf. Figure 1(b)).

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

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

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

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

3.2 Worker Interfaces

Previous works have shown the positive effects of using conversational worker interfaces [8, 51, 57]. Addressing **RQ1**, we compare conventional web-based worker interfaces with novel conversational interfaces, not only to investigate the effect of using avatars in traditional microtask crowdsourcing, but also to study whether the use of avatars can have additional benefits for conversational interfaces.

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

To investigate whether the use of avatars can further improve the effectiveness of conversational interfaces, we developed a tool named TickTalkTurk[5] [59] for deploying text-based conversational crowdsourced microtasks on popular crowdsourcing platforms. The conversational interface deployed by this tool is also built using HTML, CSS, and Javascript, and therefore compatible with most crowdsourcing platforms. We refer to the conversational interfaces as **Chat** in figures and tables henceforth in this paper. On the conversational interface, the task instructions, avatar customization, microtasks, and surveys are sent to workers via messages, from a gender-neutral conversational agent named "Andrea" with the profile image of a droid. Workers then reply to

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

Fig. 2. Worker interfaces for microtask crowdsourcing. (a), (b), and (c) represent conventional web worker interfaces (Web). (d), (e), and (f) represent novel conversational worker interfaces (Chat).

messages using a simple text field, or use the provided input elements (i.e., buttons, sliders) to respond to questions and tasks presented by the conversational agent.

Based on the task types being served (cf. Section 3.3), we provide three input types: *1) Single-selection*: this input type is used for workers to select one answer from multiple choices, which is implemented using `radio buttons` and `customized buttons` respectively on Web and Chat interfaces; *2) Free-text*: this input type is used for providing open-ended answers. Workers are required to input their answers via a `textarea` HTML element on the Web interface, or type their answers and send to the conversational agent as `messages` on the Chat interface; *3) Slider*: workers can move a handle to indicate a value on the slider. Both Web and Chat interfaces use HTML-based `slider` elements to provide input for some specific types of questions.

During task execution in the conventional Web interfaces, the customized avatar (either with or without the `characterization label`) is displayed on the left side of the input element that the worker is focusing on. We chose to position the avatar visibly to ensure that workers can always see the avatar and have opportunities to identify with their customized avatars, as shown in Figure 2 (b) and (c). Similarly, on the conversational Chat interface, the customized avatar is always displayed instead of the users' profile image, as shown in Figure 2 (e) and (f).

3.3 Microtask Design

We chose the task types of Image Transcription and Information Finding to conduct our experiments, and investigated the impact of using avatar-related affordances on task performance across these two task types. These two task types are popularly crowdsourced and have been widely used in crowdsourcing research [18, 23, 57]. Image Transcription tasks are relatively easy but can be highly monotonous. Information Finding tasks are relatively difficult, but workers can gain new knowledge while searching the web for relevant meta-data during task execution. Previous work has shown that conversational interfaces can employ text input as an alternative to other types of input. For example, multiple-selection can be realized by asking users to type option labels/numbers [51]. For the scope of our work in this paper, we only consider textual input as the input type for the tasks. A variety of input types, such as multiple-selection, sliders, or even bounding boxes, can be studied in the imminent future.

3.3.1 Image Transcription. In these tasks, workers view the images randomly generated by Claptcha[6] and transcribe the text displayed in the images. By using Claptcha, the actual text in the image, the image size, and the strength of noise can be easily tuned. The images for transcription are automatically generated, containing 5 - 18 random, distorted English letters (upper case or lower case) with Gaussian white noise. Image Transcription microtasks need relatively less time and effort compared to the Information Finding tasks described below.

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

The celebrities in the list are selected to represent different level of complexity. For instance, finding the middle name of Alan Turing is not ambiguous, while the name of computer scientist Michael Jordan will also show results for the famous basketball player Michael Jordan. Compared to Image Transcription, each Information Finding microtask needs more time, but workers have an opportunity to gain new knowledge (e.g. to learn about more famous people and some potentially interesting facts) while completing the tasks.

3.4 Measures

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

Perceived Workload. We use NASA's Task Load IndeX (NASA-TLX) questionnaire[7] to measure workers' perceived workload. The NASA-TLX questionnaire evaluates worker's cognitive workload while completing tasks on six dimensions — Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Mental Demand and Physical Demand can measure how mentally or physically demanding the crowdsourcing task was. Temporal Demand can be interpreted as how hurried or rushed the pace of task execution was. Performance and Effort represent how successful the performance was and how hard the task was respectively, while accomplishing the task. Finally, Frustration indicates how stressed and annoyed the workers felt during task execution.

[6] https://github.com/kuszaj/claptcha
[7] https://humansystems.arc.nasa.gov/groups/TLX/

Workers are required to indicate their feelings on each dimension using a slider ranging from 0 to 20. The TLX scores are later scaled to 0 to 100. The lower the TLX score is, the less mental demand, less physical demand, less temporal demand, more successful performance, less effort, and less frustration are perceived by the worker.

Intrinsic Motivation. We use the Intrinsic Motivation Inventory (IMI) [52] to measure worker's intrinsic motivation to understand whether workers enjoy using the avatars, and thus how motivated during task execution they are. IMI has been widely used to assess play experience, and prior work has shown that self-identification with avatars can increase the intrinsic motivation of players [5, 27].

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

Worker Accuracy. We use the percentage of correctly answered microtasks to measure worker accuracy. Specifically, in Information Finding tasks, a microtask is considered as correctly answered if and only if the answer provided by the worker contains the true middle name of the corresponding famous person, e.g., Irwin for Michael Jordan (computer scientist). In Image Transcription tasks, to maintain a reasonable task difficulty level, we added relatively strong artificial noises (Gaussian white noises, $0.1 \leq \sigma \leq 0.5$) and distortions into the images using Claptcha. This results in some completely illegible letters (roughly around 20% on manual inspection by the authors). Therefore, we use one of the most common string similarity metrics - the Levenshtein distance to measure the difference between the answer and the expected value [45]. In this work, we thereby tolerate 20% of mismatches. Thus, the answer for an image transcription microtask is considered to be correct if and only if the Levenshtein similarity ratio between the answer provided by workers and the expected value is greater than 80%. The Levenshtein similarity ratio is calculated as:

$$\textit{Levenshtein similarity ratio} = \frac{|a| + |b| - lev(a,b)}{|a| + |b|}, \tag{1}$$

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

Worker Retention. We use the number of answered optional microtasks to measure worker retention. For each worker, there are at most 50 available microtasks (including mandatory microtasks and optional microtasks). As described earlier, workers first have to answer 5 mandatory microtasks, ensuring that we collect sufficient data for analyzing worker performance in terms of their accuracy and execution time. Workers cannot submit the answers if the 5 mandatory microtasks are not completed. After that, workers can complete as many of the 45 optional microtasks to follow as they wish.

4 EXPERIMENTS

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

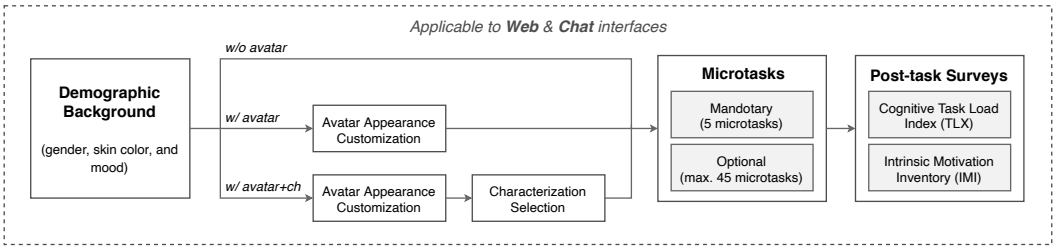


Fig. 3. Experimental procedure with three avatar conditions (*w/o avatar*, *w/ avatar*, and *w/ avatar+ch*).

4.1 Experimental Conditions

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

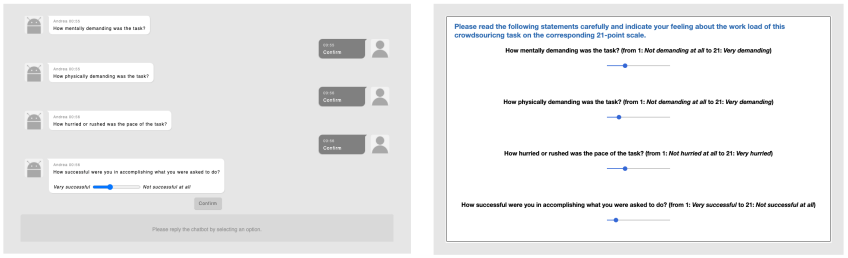
4.2 Procedure

The experiment is performed following the procedure displayed in Figure 3. Workers are required to first answer a few questions about their backgrounds. Before executing the crowdsourcing microtasks, workers in avatar-related experimental conditions will be guided through avatar customization. After executing five mandatory microtasks, workers can complete as many of the 45 optional microtasks as they wish to. Finally, workers are asked to complete two post-task surveys corresponding to their perceived workload and intrinsic motivation respectively. The details of the experimental procedure are explained below.

4.2.1 Demographic Background. The objectives of asking demographic background questions are 1) to understand the demographic distribution of the workers, and 2) to initialize the avatar appearance according to workers' background, as we described in Section 3.1. During this step, we ask workers three questions about gender, skin color, and mood respectively. There are four available gender options (non-binary, female, male, and others), seven available skin colors (tanned, yellow, pale, light, brown, dark brown, and black), and nine types of moods in three main categories (pleasant, unpleasant, and neutral). The instrument for measuring mood is called Pick-A-Mood [15], which is a robust and validated tool, and has been used in multiple HCI studies [58, 68].

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

Workers assigned to the **w/o avatar** condition are directly asked to complete the microtasks (5 mandatory, 45 optional) after responding to the demographic background questions. While completing the microtasks in this condition, workers do not have a corresponding avatar, as shown in Figure 2 (a) and (d).



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

Fig. 4. User interfaces for workers to complete the post-task survey.

Workers in the **w/ avatar** condition have an opportunity to customize the visual appearance of their avatars. On completing the customization of their avatar's appearance, workers are asked to complete the microtasks (5 mandatory, 45 optional). Thus, a customized avatar is displayed throughout task execution, as shown in Figure 2 (b) and (e).

In the **w/ avatar+ch** condition, workers are required to do proceed through avatar appearance customization and avatar characterization selection, before they can complete the microtasks (5 mandatory, 45 optional). Therefore, the customized avatars are displayed while workers complete the microtasks, along with a `characterization label` below the avatars in each case, as shown in Figure 2 (c) and (f).

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

4.2.4 Post-task Surveys. After completing the microtasks, workers are asked to complete two questionnaires. The first survey is the NASA-TLX for measuring workers' perceived workload. On both conversational interface (Chat) and conventional web interface (Web), workers should provide their answers using `slider` elements. In this study, workers in Chat conditions use the Chat interface and workers in Web conditions use the Web interface to complete post-task surveys. We did not redirect the workers in Chat conditions back to conventional web-based post-task surveys. This was motivated by prior work which has shown that dramatically changing UIs may affect users' mental models [65]. Furthermore, previous work has shown that a conversational interface does not affect the actual results [57]. Therefore, the results from our post-task surveys are reliable and valid. The second survey is a subset of the Intrinsic Motivation Inventory (IMI) for measuring workers' enjoyment and effort exerted during task execution.

The post-task survey is implemented on both Chat and Web interfaces. Survey questions on both interfaces are exactly the same as the original metrics. A screenshot of survey interfaces is

shown in Figure 4. Workers are expected to input the answer by using sliders/customized buttons in conversational interfaces, and sliders/radio buttons in conventional web interfaces respectively.

4.3 Cost and Quality Control

We recruited participants from the Amazon's Mechanical Turk (AMT) crowdsourcing platform. We set up 3 avatar conditions (**w/o avatar**, **w/ avatar**, and **w/ avatar+ch**) and 2 interfaces (**Web**, and **Chat**), resulting in $3 \times 2 = 6$ experimental conditions. For each condition, we published 60 Human Intelligence Tasks (HITs), and each HIT is completed by a unique worker following the between-subjects experimental procedure described in Section 4.2.

In order to avoid learning biases, each worker could complete only a single HIT throughout our entire experiment. To ensure this, we stored each worker's unique AMT `WorkerID`. If a WorkerID was already recorded in our database, the task content was not rendered, and the corresponding worker was kindly informed to exit the HIT. In total, we recruited 60 unique workers per condition. In each condition, we randomly distributed the 60 workers into the two task types evenly (30 unique workers per task type). Thus, $60 \times 6 = 360$ unique workers participated in our experiment. To further ensure reliable participation, we used a qualification type provided by AMT — each worker's overall HIT acceptance rate had to be greater than 95%. In addition, a worker who has one of the following behaviors is regarded as a malicious worker [24]: 1) accuracy is 0 and entering the same answer for all the questions; 2) accuracy is 0 and always entering meaningless random strings (not words). Therefore, we manually inspected the crowdsourced results and excluded 8 workers who exhibited obvious unreliable behavior. The excluded workers were not replaced in this study since they only account 2% of the total size..

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

5 RESULTS

5.1 Demographic Distribution

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

5.2 Perceived Workload

The mean TLX scores for all six dimensions are illustrated in Figure 5. According to normality tests, the TLX score distributions come from a normal distribution (the average skewness is 0.2, the average kurtosis is -2.0, the average Shapiro-Wilk statistic score W is 0.88). To see if significant differences exist across the three avatar conditions (w/o avatar, w/ avatar, and w/ avatar+ch) and two interfaces (Chat and Web), we conducted two-way factorial multivariate ANOVA tests ($\alpha = 0.05$, Type I), with the null hypothesis that the mean value is the same across all six conditions

(Web w/o avatar, Web w/ avatar, Web w/ avatar+ch, Chat w/o avatar, Chat w/ avatar, and Chat w/ avatar+ch). The results of the tests are shown in Table 1. For the Image Transcription tasks, we found that worker interfaces have a significant effect on the Performance dimension, showing that a conversational interface can significantly improve the sense of success with respect to performance. For the Information Finding tasks, we found that worker interfaces have significant effects on Performance, Effort, and the overall TLX score, suggesting that a conversational interface can reduce the perceived workload of workers. Furthermore, we found that conditions with avatars have a significant effect on the dimension of Effort, showing that avatar customization, either on Web or Chat interface, can significantly reduce the perceived task difficulty. We also observe a weak effect (not significant, $p = 0.067$) of the interaction of worker interfaces and avatar conditions in Physical Demand dimension.

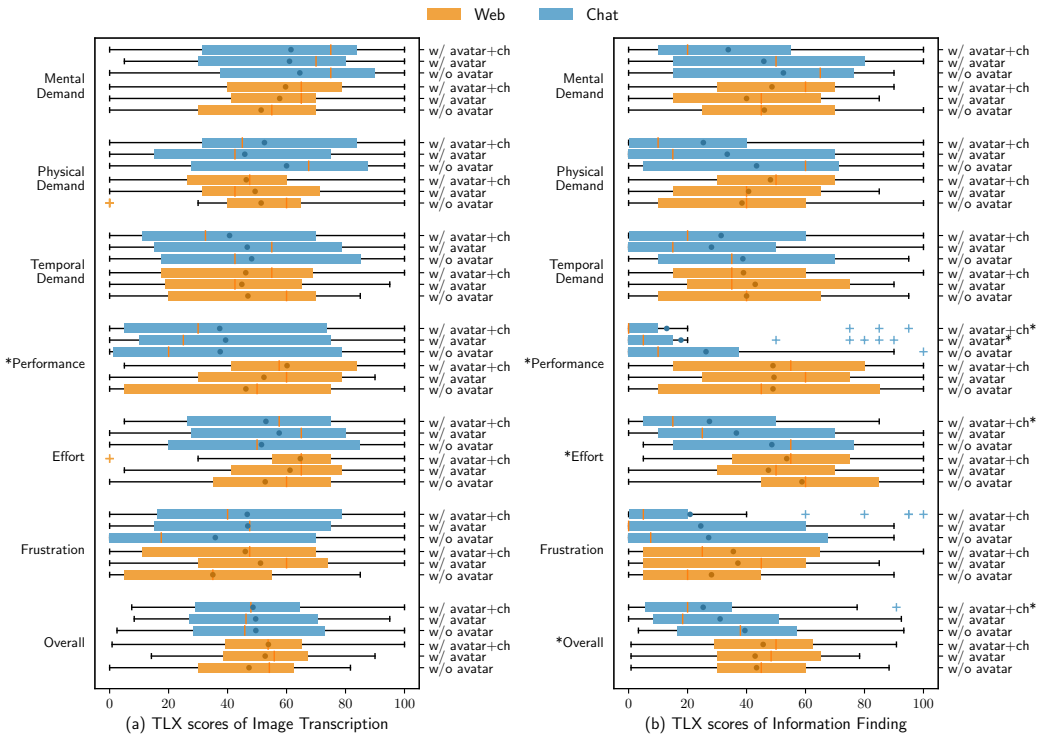


Fig. 5. Boxplots of self-reported TLX scores corresponding to (a) *Image Transcription* and (b) *Information Finding* tasks across six dimensions. Dark points represent mean values and red lines (|) represent medians. The lower the TLX score is, the less mental demand, less physical demand, less temporal demand, more successful performance, less effort, and less frustration are perceived by workers respectively. Note that the asterisk (*) on a dimension indicates a statistically significant difference between conditions resulting from an ANOVA test; the asterisk (*) on a condition indicates a statistically significant difference in comparison with the baseline condition (web w/o avatar).

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

Table 1. Results of two-way multivariate ANOVA tests (Type I) for TLX dimensions and two-way ANOVA tests (Type I) for overall TLX scores.

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

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

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

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

To interpret our data beyond p-values and better understand effect sizes in terms of the overall TLX scores, we leverage estimation plots [31], as shown in Figure 6 (the estimation plots of other TLX dimensions can be found in the companion webpage[8]). Jitter plots show all the overall TLX scores, and how they distribute, across experimental conditions. Here, we use the baseline condition – the Web interface without worker avatars (the state of the art), as a control group in the plots, to make comparison with all the other experimental conditions. The estimation plots also show the resampling distribution of the difference in means, representing the effect size. We found that the effect sizes in Image Transcription tasks were minor. However, it is still obvious that in jitter

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

plots (swarm plots), the points corresponding to conversational interfaces tend to distribute below 50 (the middle point of TLX scale), while the points corresponding to traditional Web interfaces tend to distribute above 50. In terms of Information Finding, in comparison with the baseline (Web w/o avatar), the effect sizes of worker avatars on conversational interfaces (both Chat w/ avatar and Chat w/ avatar+ch) are large, showing a possible positive impact of the interaction effect of conversational interface and worker avatar on perceived workload.

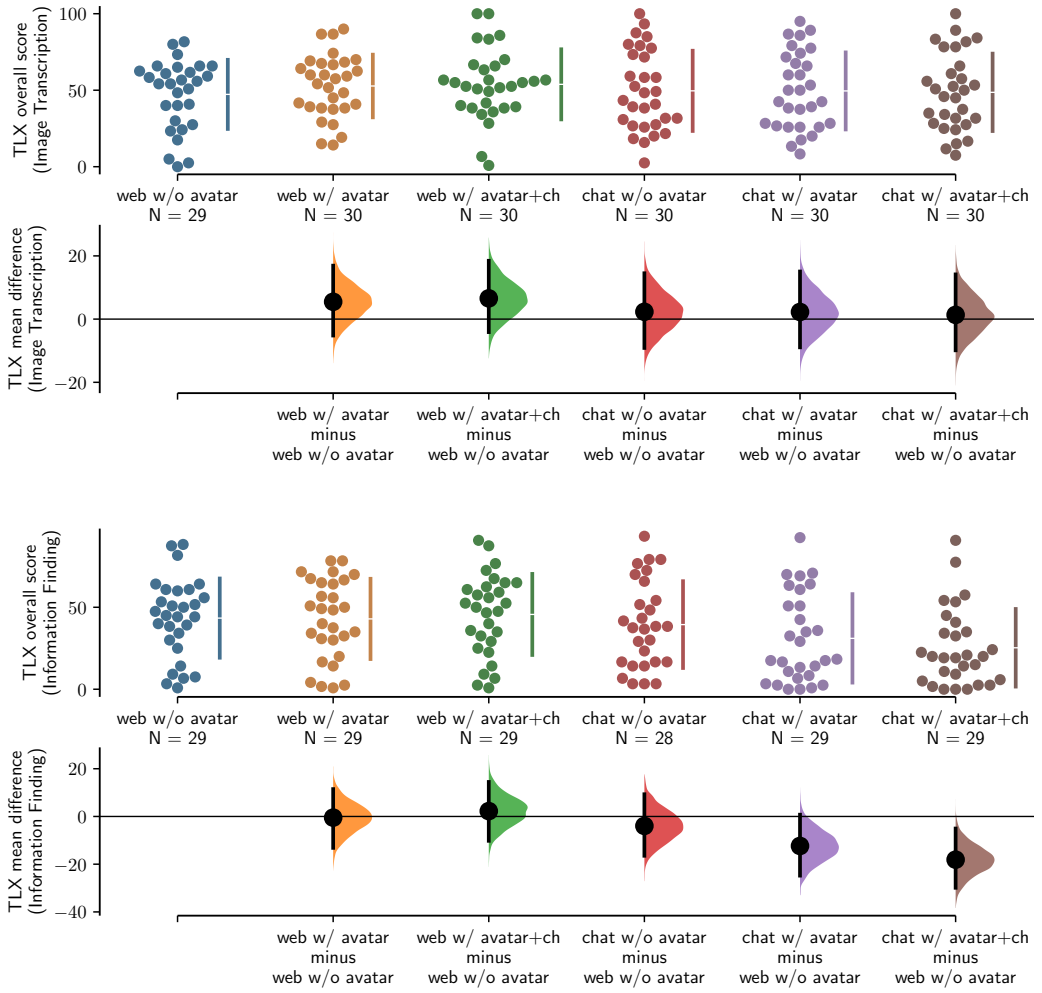


Fig. 6. Estimation plots of TLX scores of Image Transcription and Information Finding tasks.

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

5.3 Intrinsic Motivation

Figure 7 shows the IMI scores of different avatar conditions across two interfaces (Web and Chat) and two task types (Image Transcription and Information Finding), in two intrinsic motivation dimensions — Interest-Enjoyment and Effort-Importance respectively.

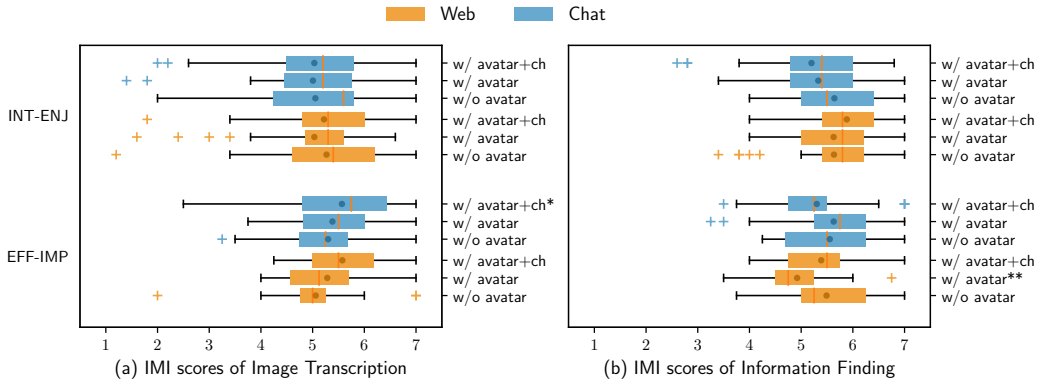


Fig. 7. Boxplots of self-reported intrinsic motivation inventory score of *Image Transcription* tasks and *Information Finding* tasks in interest-enjoyment and effort-importance dimensions, where dark points represent mean values and red lines (|) represent medians. Note that ** indicates statistical significance with Bonferroni correction and * indicates $p < 0.05$ but not significant after Bonferroni correction).

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

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

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

5.4 Objective Worker Performance

5.4.1 Worker Retention. The results of worker retention, measured by the number of answered optional questions, are shown in Figure 8 (a) and (b). According to normality tests, the worker retention does not follow a normal distribution (the average skewness is 4.3, the average kurtosis is 3.1, the average Shapiro-Wilk statistic score is 0.57). Therefore, we used Mann-Whitney U tests to find differences in worker retention across conditions measured by the number of answered optional microtasks. The results are in-line with our recent findings [57]. The conversational interfaces (Chat) were found to be more effective in retaining workers in both Image Transcription and Information Finding tasks, compared to Web interfaces ($p = 0.026$, CL effect size $f = 0.57$, and $p = 0.085$, CL effect size $f = 0.56$ respectively).

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

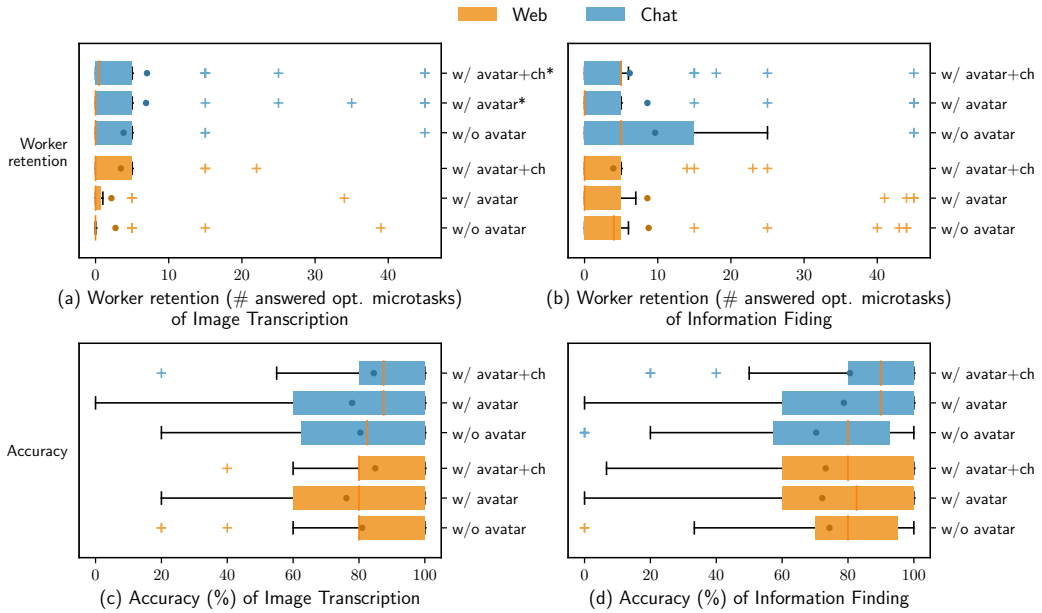


Fig. 8. Boxplots of worker retention measured by the number of answered optional microtasks, and worker accuracy (%) measured by the percentage of correctly answered microtasks, corresponding to *Image Transcription* tasks and *Information Finding* tasks. Dark points represent mean values, red lines (|) represent medians, and (* represents significant difference in comparison with the baseline).

5.4.2 Worker Accuracy. Results pertaining to worker accuracy are shown in Figure 8 (c) and (d). Aligned with previous work [57], we found no significant difference between experimental conditions across the two task types — Image Transcription ($p > 0.18$) and Information Finding ($p > 0.1$), according to Mann-Whitney U tests (since worker accuracy does not come from normal distributions as per normality tests: the average skewness is -2.7; the average kurtosis is 1.2; the average Shapiro-Wilk statistic score is 0.81). However, as shown in Table 2, we found that the condition with avatar appearance customization and characterization selection (w/ avatar+ch)

corresponds the highest worker accuracy in three out of four cases (Image Transcription on both Web and Chat interfaces, and Information Finding on Chat interface). Apart from our observation in the Information Finding tasks on the Web interface (where all three avatar conditions correspond to similar worker accuracy with only 1-2% differences), the mean values of worker accuracy of the avatar appearance customization and characterization selection condition (w/ avatar+ch) are 5%-13% higher than the baseline condition (w/o avatar).

Table 2. Worker accuracy (unit:%, $\mu \pm \sigma$) measured by the percetage of correctly answered microtasks, where the highest values among each interface are displayed in bold.

	Image Transcription		**Information Finding**	
Condition	Web	Chat	Web	Chat
w/o avatar	80 ± 21	80 ± 22	**74 ± 27**	70 ± 31
w/ avatar	76 ± 22	77 ± 23	72 ± 32	78 ± 25
w/ avatar+ch	**84 ± 15**	**84 ± 17**	73 ± 30	**80 ± 23**

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

5.5 Avatar Appearance Customization and Characterization Selection

5.5.1 Avatar Customization Time. In terms of the time spent on appearance customization, workers in the Web interface conditions (29.66 ± 31.76 seconds) spent slightly longer time on customizing avatars in comparison with workers in the Chat interface conditions (23.35 ± 31.39 seconds). To analyze the impact of customization time on worker performance, we split the workers into three groups according to the standard deviation ($\pm 0.5\sigma$) of avatar customization time, resulting in — a group of workers with short customization time (customization time $< \mu - 0.5\sigma$, less than 10.6 seconds), a group of workers with medium customization time ($\mu - 0.5\sigma \leq$ customization time $< \mu + 0.5\sigma$, 10.6-42.3 seconds), and a group of workers with long customization time (customization time $\geq \mu + 0.5\sigma$, longer than 42.3 seconds). As shown in Table 3, we found that the group of workers corresponding to a long customization time exhibit the highest worker accuracy (83 ± 22%) in comparison with the group of workers with short customization time (accuracy = 77 ± 27%, $p = 0.065$, CL effect size $f = 0.57$), and with medium customization time (accuracy = 77 ± 24%, $p = 0.021$, CL effect size $f = 0.60$) using Mann-Whitney U tests. The results suggest that workers spending a longer time on avatar customization go on to perform with a higher accuracy in general. This may be explained by a greater level of self-identification through avatar customization which leads to an increased intrinsic motivation, as supported by our findings.

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

For all the workers in avatar conditions, we found that on average, the avatar customization time (26.48 ± 31.73 seconds) occupies 8.14% (±10.01%) of the total task execution time (426.03 ± 399.73

Table 3. Worker accuracy (unit: %, $\mu \pm \sigma$) and execution time per microtask (unit: second, $\mu \pm \sigma$) of three groups divided according to avatar customization time.

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

seconds). From the perspective of the task requester, facilitating avatar customization may not appear to be a useful investment for short or less complex batches of tasks, considering the additional costs that requesters may incur in return for limited positive effects. However, for long, complex, or challenging tasks, facilitating avatar customization can warrant the reasonable overheads with an aim to effectively improve worker experience. We envision that in the future, avatar customization can be a feature that is supported by crowdsourcing platforms rather than by individual task requesters with an aim to foster a healthy work experience for crowd workers.

5.5.2 Self-Identification with Worker Avatars. We explored the number of workers who actually changed the appearance of their initial avatar, that was generated with 3 parameters in accordance to their demographic backgrounds. We found that only 58 workers (24%) changed their skin colors for their avatars, while most of these workers (37 out of 58) just slightly tuned the skin color (for instance, change between Black and Dark Brown, Dark Brown and Brown, Brown and Light, or Light and Pale), suggesting that most workers were generally satisfied with their initialized avatars based on the demographic information they provided. We found that 26 workers (11%) changed the depicted moods for their avatars. 19 out of 26 workers changed their moods from either pleasant or unpleasant to neutral; only 4 workers changed to pleasant moods from unpleasant moods, while 3 workers did the reverse. As for accessories and clothing colors, we found that all the types of accessories are nearly equally distributed, and more workers chose Black for avatar's clothing while the least number of workers chose Yellow (50 Black and 29 Yellow, average = 40). Our findings pertaining to avatar customization indicate that workers generally cared about the appearance of their avatars, and this suggests the potential emergence of self-identification [5].

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

Diligent workers are described as workers who exhibit a high accuracy but a relatively slower task execution speed, and Competent workers are described as workers with a reasonably high accuracy but a faster task execution speed. As shown in Table 4, on exploring the mean worker accuracies of the workers who selected the "Competent" characterization, we found that they exhibited a higher accuracy than the workers who selected a "Diligent" characterization in both Image Transcription and Information Finding tasks (9% and 24% higher respectively). Interestingly, in terms of task execution time per microtask (speed), the workers who selected the "Diligent" characterization exhibited faster execution speeds in comparison with those workers who selected either the "Competent" or "Balanced" characterizations (7% and 15% faster respectively) in Information Finding tasks. This can be explained by workers' wishful identification with selected characterizations – workers were probably aware of their real characterizations (and shortcomings), therefore they

Table 4. Worker accuracy (unit: %, $\mu \pm \sigma$) and execution time per microtask (unit: second, $\mu \pm \sigma$) of *Image Transcription* tasks and *Information Finding* tasks across *Diligent*, *Competent*, and *Balanced* characterizations.

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

may have chosen avatar characterizations which they aspired to. This is consistent with what has been observed in gaming research [5].

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

6 DISCUSSION

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

By analyzing the results about avatar appearance customization and characterization selection, we found that customization of the avatar's appearance facilitates similarity identification among workers, while the avatar characterization facilitates wishful identification among workers. Our results show that **58**% of workers selected the "Balanced" characterization. Furthermore, we found that the performance of workers, to a large extent, does not follow the avatar characterization they selected. For example, **28**% of workers who selected a Diligent characterization ended up performing with a relatively low accuracy, **14**% of workers who selected a Competent characterization ended up exhibiting relatively long task execution times. This can potentially be explained by the emergence of wishful identification with avatar characterizations, and these workers falling short of their aspirations (i.e., to complete tasks with higher accuracies or lower task execution times respectively). Additional experiments are required to distill the extent to which avatar appearance customization, and avatar characterization in addition to appearance customization systematically facilitate similarity and wishful identification. Nevertheless, our findings bear evidence to support that avatar identification can be an effective tool for improving satisfaction and enjoyment during crowdsourcing task execution [39, 66].

6.1 Using Avatars in Conversational Interfaces to Improve Crowdsourcing

Conversational interfaces for microtask crowdsourcing have emerged owing to the concomitant advantages of better engaging users. Our study has revealed that using avatars in conversational interfaces can ease the perceived workload of workers, by improving the sense of success in performance and reducing the perceived task difficulty while completing tasks. Specifically for tasks that are more challenging, we found that conversational interfaces with avatar appearance

customization and avatar characterization selection can be effective in significantly reducing cognitive workload from the perspectives of performance (75% lower TLX score), and effort (53% lower TLX score), in comparison to the baseline of conventional web interfaces without worker avatars.

6.2 Implications for Design

Alleviating perceived workload. Our results show that the conversational interface with the functionality of avatar appearance customization and characterization selection can effectively decrease cognitive workload. Particularly, the conversational interface with worker avatars can significantly make workers feel more successful, and perceive less difficulty, while completing tasks. This finding has important implications in future crowdsourcing task design. The mental state of crowd workers has become a major concern due to an increase in the number of workers who work full-time and earn their livings in crowdsourcing marketplaces, coupled with power asymmetry and other challenges workers typically face [26, 35, 49, 64]. Most state-of-the-art tools and approaches in the field of crowdsourcing are developed from the perspective of task requesters. Although human factors and worker-friendly interventions have been considered, the focus has largely been on improving the quality of outcomes [37, 40], rather than ensuring the wellbeing and the mental health of workers. The avatar customization framework we introduce in this work is developed completely based on HTML/CSS/Javascript, and is designed to be compatible and portable. Using the avatar framework we designed and made publicly available, there is very little overhead involved in integrating the use of worker avatars in microtask crowdsourcing platforms – task requesters can readily integrate avatar customization into their tasks. In exchange for the small overhead of integrating avatar customization, task requesters can reap worthy benefits of reducing the perceived task difficulty among workers and increase their sense of success.

Facilitating avatar identification. Avatar identification in crowdsourcing can be interpreted as the resonance and identification of crowd workers with an avatar that represents them. Avatar identification has been shown to be useful for fostering intrinsic motivation, increasing satisfaction and entertainment, and improving preventive health outcomes [5, 39, 66]. Our results pertaining to avatar appearance customization and characterization selection imply that similarity identification and wishful identification can be facilitated among workers in microtask crowdsourcing to reap similar rewards.

Strengthening the relationship between task requesters and workers. A strong and healthy relationship between task requesters and workers is crucial to all relevant stakeholders in the crowdsourcing paradigm. Maintaining a good relationship can assist task requesters in building their reputation and attracting more workers of high quality. For crowd workers, a good relationship can help them maintain credible profiles, increase their access to more work in the crowdsourcing marketplaces, maximize their earnings, and reduce the emotional toll and frustration that can result from mistrust and rejection [54]. By alleviating the perceived workload and improving the sense of success among workers through the use of avatars and conversational interfaces, there lies a great potential to further foster healthy requester-worker relationships. This bears a useful implication on ensuring the sustainability of crowdsourcing marketplaces.

6.3 The Value of Using Worker Avatars in Crowdsourcing

We note that conversational interfaces reduce the perceived cognitive load of workers in comparison to conventional web interfaces, corroborating recent findings. Critically reflecting on our collective findings, the added value of using worker avatars is less prominent in relatively easy tasks such as solving CAPTCHAs (Image Transcription), and along the dimensions of Mental Demand, Physical

Demand, Temporal Demand, and Frustration. We found that using worker avatars in relatively more difficult tasks (Information Finding), led to a reduction in the perceived cognitive load of workers in conversational interfaces. Our findings hint that worker avatars can play a more significant role in tasks that are relatively more difficult, but include elements of learning (as indicated by the open-ended comments from workers). Another explanation for our null findings with respect to the impact of avatar customization in image transcription tasks can be that due to the relatively less amount of time required for task completion, workers do not meaningfully self-identify with their avatars. Future experiments can explore how self-identification with avatars is mediated by the task execution time.

6.4 Limitations and Future Work

Previous studies about motivations in gaming systems have shown the effectiveness of fostering intrinsic motivation of the player by using avatar customization [5]. However, critically reflecting on our findings in microtask crowdsourcing, we found that avatar appearance customization had no significantly positive impact on the intrinsic motivation of crowd workers (with respect to interest-enjoyment and effort-importance). This can potentially be explained by the fact that workers are mainly motivated by monetary incentives in paid crowdsourcing marketplaces, rather than by the allure to stimulate their feelings of enjoyment and interest through task execution. Based on our findings in the paid microtask crowdsourcing setup, future work can explore the potential of avatar customization in voluntary crowdsourcing.

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

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

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

In contrast, workers found the Information Finding tasks to be interesting, since they provided workers with a chance to gain new knowledge and learn some interesting facts through the course of executing the tasks. Example comments reflecting these notions are shared below. It is interesting to note these perceptions despite the fact that Information Finding tasks required more effort and were time-consuming (Information Finding 64.6 ± 59.8 vs Image Transcription 35.9 ± 24.7, unit: seconds per microtask).

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

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

Our findings suggest that the task type can influence workers' experience, and using worker avatars can be more effective in tasks that are more challenging and exploratory, especially when they involve elements or opportunities for learning [19].

In this study, the post-task survey was conducted on either the Chat interface or the Web interface depending on the experimental condition. However, a previous study has pointed out that a "casual"

conversational style could influence participants' responses [38]. To avoid this potential confound, in our post-task surveys we used a "formal" conversational style (which is known to not influence worker responses). We emphasize the caveat that other complex factors may indeed affect the nature of responses when a conversational interface is employed.

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

7 CONCLUSIONS

In this work, we carried out a study to investigate whether using worker avatars and enabling avatar customization (through avatar appearance and characterization) can improve worker experience in conventional web and novel conversational interfaces. To address **RQ1**, we designed worker interfaces combining the avatar customization affordances. We carried out a between-subjects experimental study with 360 workers to analyze their perceived workload, intrinsic motivation, and the concomitant quality-related outcomes. Results suggest that using avatar appearance customization and avatar characterization selection in microtask crowdsourcing, especially combined with conversational interfaces, has positive effects on reducing workers' perceived workload and improving quality-related outcomes, which can benefit workers in terms of improving the sense of success and lowering the perceived task difficulty. To address **RQ2**, we evaluated how worker avatar appearance and characterization could affect intrinsic motivation during task execution, and studied the traits of the workers who selected different characterizations. Our results reveal that workers who put more effort into avatar customization exhibited higher accuracy. The results also show the emergence of similarity and wishful identifications. Our findings have important implications on the design of crowdsourcing tasks, and improving worker experience during task execution.

ACKNOWLEDGMENTS

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative (no. e-infra190082).

REFERENCES

[1] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Science robotics* 3, 21 (2018).

[2] Timothy Bickmore and Justine Cassell. 2001. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 396–403.

[3] Timothy Bickmore and Justine Cassell. 2005. Social dialongue with embodied conversational agents. In *Advances in natural multimodal dialogue systems*. Springer, 23–54.

[4] Timothy W Bickmore, Suzanne E Mitchell, Brian W Jack, Michael K Paasche-Orlow, Laura M Pfeifer, and Julie O'Donnell. 2010. Response to a relational agent by hospital patients with depressive symptoms. *Interacting with computers* 22, 4 (2010), 289–298.

[5] Max V Birk, Cheralyn Atkins, Jason T Bowey, and Regan L Mandryk. 2016. Fostering intrinsic motivation through avatar identification in digital games. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2982–2995.

[6] Max Valentin Birk and Regan Lee Mandryk. 2019. Improving the efficacy of cognitive training for digital mental health interventions through avatar customization: crowdsourced quasi-experimental study. *Journal of medical Internet research* 21, 1 (2019), e10133.

[7] Anne Bowser, Derek Hansen, Yurong He, Carol Boston, Matthew Reid, Logan Gunnell, and Jennifer Preece. 2013. Using gamification to inspire new citizen science volunteers. In *Proceedings of the first international conference on gameful design, research, and applications*. 18–25.

[8] Luka Bradeško, Michael Witbrock, Janez Starc, Zala Herga, Marko Grobelnik, and Dunja Mladenić. 2017. Curious Cat–Mobile, Context-Aware Conversational Crowdsourcing Knowledge Acquisition. *ACM Transactions on Information Systems (TOIS)* 35, 4 (2017), 1–46.

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

[10] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, et al. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (2010), 756–760.

[11] Nicole Crenshaw and Bonnie Nardi. 2014. What's in a name? Naming practices in online video games. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play*. 67–76.

[12] Mihaly Csikszentmihalyi and Mihaly Csikzentmihaly. 1990. *Flow: The psychology of optimal experience*. Vol. 1990. Harper & Row New York.

[13] 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 Computing Surveys (CSUR)* 51, 1 (2018), 1–40.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

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

[16] Laurence Devillers, Sophie Rosset, Guillaume Dubuisson Duplessis, Lucile Bechade, Yucel Yemez, Bekir B Turker, Metin Sezgin, Engin Erzin, Kevin El Haddad, Stephane Dupont, et al. 2018. Multifaceted engagement in social interaction with a machine: The joker project. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 697–701.

[17] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of mechanical Turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 135–143.

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

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

[20] Carsten Eickhoff, Christopher G Harris, Arjen P de Vries, and Padmini Srinivasan. 2012. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 871–880.

[21] Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2019. Crowd anatomy beyond the good and bad: Behavioral traces for crowd worker modeling and pre-selection. *Computer Supported Cooperative Work (CSCW)* 28, 5 (2019), 815–841.

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

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

[24] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1631–1640.

[25] Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2017. Towards Designing Cooperative and Social Conversational Agents for Customer Service.. In *ICIS*.

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

[27] Tobias Greitemeyer and Silvia Osswald. 2010. Effects of prosocial video games on prosocial behavior. *Journal of personality and social psychology* 98, 2 (2010), 211.

[28] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. The impact of task abandonment in crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering* (2019).

[29] DANULA HETTIACHCHI, NIELS VAN BERKEL, VASSILIS KOSTAKOS, and JORGE GONCALVES. 2020. CrowdCog: A Cognitive Skill based System for Heterogeneous Task Assignment and Recommendation in Crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 4 (2020).
[30] E Tory Higgins. 1987. Self-discrepancy: a theory relating self and affect. *Psychological review* 94, 3 (1987), 319.
[31] Joses Ho, Tayfun Tumkaya, Sameer Aryal, Hyungwon Choi, and Adam Claridge-Chang. 2019. Moving beyond P values: data analysis with estimation graphics. *Nature methods* 16, 7 (2019), 565–566.
[32] Cynthia Hoffner. 1996. Children's wishful identification and parasocial interaction with favorite television characters. *Journal of Broadcasting & Electronic Media* 40, 3 (1996), 389–402.
[33] Cynthia Hoffner and Martha Buchanan. 2005. Young adults' wishful identification with television characters: The role of perceived similarity and character attributes. *Media psychology* 7, 4 (2005), 325–351.
[34] Ting-Hao Kenneth Huang, Joseph Chee Chang, and Jeffrey P Bigham. 2018. Evorus: A Crowd-powered Conversational Assistant Built to Automate Itself Over Time. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 295.
[35] Lilly C Irani and M Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 611–620.
[36] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2011. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. 1941–1944.
[37] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2013. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information retrieval* 16, 2 (2013), 138–178.
[38] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.
[39] Youjeong Kim and S Shyam Sundar. 2012. Visualizing ideal self vs. actual self through avatars: Impact on preventive health outcomes. *Computers in Human Behavior* 28, 4 (2012), 1356–1364.
[40] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1301–1318.
[41] 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*. 1–8.
[42] 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*. 151–162.
[43] Joey J Lee, Pinar Ceyhan, William Jordan-Cooley, and Woonhee Sung. 2013. GREENIFY: A real-world action game for climate change education. *Simulation & Gaming* 44, 2-3 (2013), 349–365.
[44] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2020. Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–27.
[45] Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. Soviet Union, 707–710.
[46] Q Vera Liao, Werner Geyer, Michael Muller, and Yasaman Khazaen. 2020. Conversational Interfaces for Information Search. In *Understanding and Improving Information Search*. Springer, 267–287.
[47] Ian J Livingston, Carl Gutwin, Regan L Mandryk, and Max Birk. 2014. How players value their characters in world of warcraft. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1333–1343.
[48] Andrew Mao, Ece Kamar, and Eric Horvitz. 2013. Why stop now? predicting worker engagement in online crowdsourcing.. In *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing*. AAAI, 103–111.
[49] David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 224–235.
[50] Elaine Massung, David Coyle, Kirsten F Cater, Marc Jay, and Chris Preist. 2013. Using crowdsourcing to support pro-environmental community activism. In *Proceedings of the SIGCHI Conference on human factors in Computing systems*. 371–380.
[51] Panagiotis Mavridis, Owen Huang, Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2019. Chatterbox: Conversational interfaces for microtask crowdsourcing. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. 243–251.

[52] Edward McAuley, Terry Duncan, and Vance V Tammen. 1989. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research quarterly for exercise and sport* 60, 1 (1989), 48–58.

[53] Michael P McCreery, S Kathleen Krach, Peter G Schrader, and Randy Boone. 2012. Defining the virtual self: Personality, behavior, and the psychology of embodiment. *Computers in Human Behavior* 28, 3 (2012), 976–983.

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

[55] Benedikt Morschheuser, Juho Hamari, Jonna Koivisto, and Alexander Maedche. 2017. Gamified crowdsourcing: Conceptualization, literature review, and future agenda. *International Journal of Human-Computer Studies* 106 (2017), 26–43.

[56] Carman Neustaedter and Elena A Fedorovskaya. 2009. Presenting identity in a virtual world through avatar appearances.. In *Graphics Interface*. 183–190.

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

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

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

[60] Albert Rizzo, Russell Shilling, Eric Forbell, Stefan Scherer, Jonathan Gratch, and Louis-Philippe Morency. 2016. Autonomous virtual human agents for healthcare information support and clinical interviewing. In *Artificial intelligence in behavioral and mental health care*. Elsevier, 53–79.

[61] Markus Rokicki, Sergej Zerr, and Stefan Siersdorfer. 2015. Groupsourcing: Team competition designs for crowdsourcing. In *Proceedings of the 24th international conference on world wide web*. 906–915.

[62] Richard M Ryan, C Scott Rigby, and Andrew Przybylski. 2006. The motivational pull of video games: A self-determination theory approach. *Motivation and emotion* 30, 4 (2006), 344–360.

[63] Martin Saerbeck, Tom Schut, Christoph Bartneck, and Maddy D Janse. 2010. Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1613–1622.

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

[65] John W Satzinger and Lorne Olfman. 1998. User interface consistency across end-user applications: the effects on mental models. *Journal of Management Information Systems* 14, 4 (1998), 167–193.

[66] Sabine Trepte and Leonard Reinecke. 2010. Avatar creation and video game enjoyment. *Journal of Media Psychology* (2010).

[67] Peter Vorderer, Christoph Klimmt, and Ute Ritterfeld. 2004. Enjoyment: At the heart of media entertainment. *Communication theory* 14, 4 (2004), 388–408.

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