TU Delft

Delft University of Technology

“Are we all in the same boat?” Customizable and Evolving Avatars to Improve Worker Engagement and Foster a Sense of Community in Online Crowd Work

de Groot, Esra Cemre Su; Gadiraju, Ujwal

DOI
10.1145/3613904.3642429

Publication date
2024

Document Version
Final published version

Published in
CHI 2024 - Proceedings of the 2024 CHI Conference on Human Factors in Computing Sytems

Citation (APA)
de Groot, E. C. S., & Gadiraju, U. (2024). “Are we all in the same boat?” Customizable and Evolving Avatars to Improve Worker Engagement and Foster a Sense of Community in Online Crowd Work. In *CHI 2024 - Proceedings of the 2024 CHI Conference on Human Factors in Computing Sytems* Article 640 (Conference on Human Factors in Computing Systems - Proceedings). ACM. https://doi.org/10.1145/3613904.3642429

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

“Are we all in the same boat?” Customizable and Evolving Avatars to Improve Worker Engagement and Foster a Sense of Community in Online Crowd Work

Esra Cemre Su de Groot
Delft University of Technology
Delft, The Netherlands
e.c.s.degroot@tudelft.nl

Ujwal Gadiraju
Delft University of Technology
Delft, The Netherlands
u.k.gadiraju@tudelft.nl

Figure 1: This screenshot illustrates the worker community space in one of our experimental conditions (`Evolving⊕Comm`) in which workers could customize their avatars with an evolving set of features as they progressed through a batch of tasks. The community space includes the REFRESH, SIMILAR MOOD, ORDER ON LEVEL, and LOAD MORE buttons. These allow workers to (a) see avatars of a random subset of other workers who completed the same tasks, (b) see avatars of all workers who expressed the same task-related feelings after completing the same tasks, (c) order worker avatars based on the highest level that workers progressed to within the task batch, (d) load avatars of all workers who completed the same tasks. On entering the community page, a worker's own avatar is displayed in the middle, with a random subset of other worker avatars displayed surrounding the worker. All avatars are rendered with a text bubble describing their task-related feelings.

ABSTRACT

Human intelligence continues to be essential in building ground-truth data, training sets, and for evaluating a plethora of systems. The democratized and distributed nature of online crowd work — an attractive and accessible feature that has led to the proliferation of the paradigm — has also meant that crowd workers may not always feel connected to their remote peers. Despite the prevalence of collaborative crowdsourcing practices, workers on many microtask crowdsourcing platforms work on tasks individually and are seldom directly exposed to other crowd workers. In this context, improving worker engagement on microtask crowdsourcing platforms is an unsolved challenge. At the same time, fostering a sense of community among workers can improve the sustainability and working conditions in crowd work. This work aims to increase worker engagement in conversational microtask crowdsourcing by leveraging evolving avatars that workers can customize

This work is licensed under a Creative Commons Attribution International 4.0 License.

CHI '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642429

as they progress through monotonous task batches. We also aim to improve group identification in individual tasks by creating a community space where workers can share their avatars and feelings on task completion. To this end, we carried out a preregistered between-subjects controlled study ($N = 680$) spanning five experimental conditions and two task types. We found that evolving and customizable worker avatars can increase worker retention. The prospect of sharing worker avatars and task-related feelings in a community space did not consistently affect group identification. Our exploratory analysis indicated that workers who identify themselves as crowd workers experienced greater intrinsic motivation, subjective engagement, and perceived workload. Furthermore, we discuss how task differences shape the relative effectiveness of our interventions. Our findings have important theoretical and practical implications for designing conversational crowdsourcing tasks and in shaping new directions for research to improve crowd worker experiences.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Empirical studies in HCI**.

KEYWORDS

Conversational Crowdsourcing, Worker Avatars, Group Identification, Engagement, Community

ACM Reference Format:

Esra Cemre Su de Groot and Ujwal Gadiraju. 2024. "Are we all in the same boat?" Customizable and Evolving Avatars to Improve Worker Engagement and Foster a Sense of Community in Online Crowd Work. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA*. ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/3613904.3642429

1 INTRODUCTION

The need for human input on demand has steadily increased alongside the growth in the adoption of artificial intelligence (AI) and machine learning (ML) systems across all domains [28]. The foundations of many AI systems we interact with daily rely on the labor of crowd work [30]. With the availability of crowd workers on-demand [12], human intelligence tasks (HITs) can be distributed and completed at scale on crowdsourcing platforms like Amazon Mechanical Turk,[1] Prolific,[2] and Toloka.[3] Tasks range from data labeling [10], image annotation [48], and classification [82] to the creation and support of real-time healthcare applications [3, 7].

Due to the repetitiveness of HITs, tasks can be monotonous and boring, causing task rejection and drop-out [34, 60], which is problematic for both crowd workers and task requesters. Task rejection can affect the morale of crowd workers [17], and high drop-out rates result in low-quality crowd work, also affecting worker pay. Monotonous and boring work decreases the motivation of workers [9], resulting in reduced worker engagement. Furthermore, motivation is known to be an essential factor when it comes to reducing work-related stress and burnout [87]. Similarly, job satisfaction has been shown to be positively related to subjective well-being [6]. To decrease the problematic effects of monotonous and tedious tasks for crowd workers and task distributors, we need to improve worker engagement by creating better worker experiences. In the long run, this can also result in improving the quality of crowd work [97].

Although some crowdsourcing tasks require collaboration and teamwork among workers [10, 19, 55, 66], workers typically execute microtasks individually and sometimes in isolation [24, 57]. Not all workers, therefore, have the opportunity to experience a sense of community due to this, and little is typically done to increase group identification among workers. In addition to improving worker engagement during task execution, increasing a sense of community can go a long way toward creating better worker experiences. Prior work has shown that crowd workers use external forums to communicate with other crowd workers [91, 95, 96], such as Reddit HWTF, Facebook, MTurkGrind, MTurkForum, and Turkernation. However, elaborate social interventions and facilitating extensive engagement via forums are not viable solutions for all workers. While several crowd workers have been shown to communicate with other workers, many do not communicate with others and work alone [96]. In part, this may be due to workers not having time to engage in external forums as a result of other commitments not related to crowd work [1]. It is, therefore, prudent to explore whether a lightweight method that does not require extensive social engagement or exchange of private information can still help build a sense of community among workers while completing tasks individually. Through our work, we aim to address these challenges pertaining to both research and empirical gaps.

Digital avatars are known to increase identification and user experience in online multi-player video games [89], solitary educational games [40], and conversational crowdsourcing tasks [68]. Moreover, the ability to personalize the avatar by customizing its appearance further increases users' self-identification with the avatar [5]. Prior HCI research has shown a promising impact of crowd worker avatar customization within a conversational interface to reduce cognitive workload and increase worker retention [68]. However, the notion of evolving and customizable worker avatars and their effect on worker experience and task-related outcomes remains unexplored. Addressing this research gap, we propose to couple avatar evolution and customization with workers' progress in task batches.

Since digital avatars facilitate the creation of a virtual identity [5, 63, 89], we argue that a personal worker avatar can be an effective tool to increase a sense of community among the workers while protecting their privacy. Prior research found that avatar identification relates to [90] and predicts [23] group identification in online video games. Gabbiadini et al. [23] explained that when users see their avatar in the group, they imagine themselves as being part of the group. Similarly, Takano and Taka [86] found that avatar identification has a positive effect on the feeling of belonging, partially mediated by self-expression. Inspired by this prior literature, we aim to facilitate group identification by creating a community space where workers can share their personalized avatars with other crowd workers. With the worker community space, we aim to build a lightweight intervention that can be used in tasks without elements of collaboration to reflect a feeling of

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

unity [84] by placing the virtual identity of the worker among other worker avatars. As a part of customization, the facial expressions of avatars can then be used to share (task-related) feelings with other workers in a community space on task completion, as sharing feelings (affective self-disclosure) can contribute to a feeling of connection [84]. Combining the interventions of evolving avatars and group identification, we address the following research questions in our work:

RQ1 How do evolving and customizable worker avatars affect worker experience and task-related outcomes in conversational crowdsourcing?
RQ2 To what extent can the sharing of worker avatars in a community space affect the sense of group identification among crowd workers on a crowdsourcing platform?
RQ3 How does a sense of group identification, induced by a community space where customizable and evolving avatars among crowd workers can be shared, affect worker experience and task-related outcomes in conversational crowdsourcing?

By combining avatar customization, gamified avatar evolution, and creating a sense of community, we aim to improve overall worker experiences and the quality of the task outcomes. Worker experiences can be described and measured by their *perceived workload*, *intrinsic motivation*, and *subjective engagement*. Furthermore, we aim to analyze the impact of these interventions on task-related outcomes, such as retention, accuracy, and overall task execution time. To this end, we carried out a between-subjects study by recruiting workers from the Prolific crowdsourcing platform (N = 680), spanning five experimental conditions and considering two popular types of tasks (information finding and credibility analysis). We found that evolving and customizable worker avatars can increase worker retention. Although the worker community space was not successful in fostering an increased sense of group identification among crowd workers, we found that this varied across workers based on the extent to which they considered themselves as crowd workers. Workers who identify themselves as crowd workers experience a significantly greater perceived workload, intrinsic motivation, and subjective engagement. Our findings have important implications for the design of future conversational crowdsourcing tasks and for crowdsourcing platforms, with an aim to improve worker experiences and foster a sense of community. All code and data pertaining to this work can be found in the OSF repository for the benefit of the community and in the spirit of open science.[4]

2 RELATED LITERATURE AND HYPOTHESES

We position our work in the context of worker experiences in microtask crowdsourcing and literature in the realm of creating a sense of community among users. By building on existing works in these areas, we present and ground our research hypotheses.

[4]https://osf.io/yxgcz/

2.1 Worker Engagement in Microtask Crowdsourcing

A promising way to improve worker engagement in repetitive crowdsourcing tasks is to improve the worker experience through gamification. Gamification in crowdsourcing tasks often leads to an increased motivation of crowd workers, participation and throughput rates, and quality of the work [62]. Feng et al. [21] proposed a model that describes how gamification indirectly increases the intention of participation of crowd workers by an increased level of intrinsic motivation. Examples of how to incorporate (context-independent) gamification in crowdsourcing tasks are tracking scores, leaderboards, badges/achievements, and the use of increasing levels. Highlighting the importance of targeting intrinsic motivation compared to extrinsic motivation, the results of a study by Maddalena et al. [56] suggest that while a monetary incentive may increase retention, it decreases the quality of the work. Interestingly, the same study showed that while the total number of completed voluntary tasks was higher with gamification compared to no gamification (no furtherance incentives), this effect was caused by a number of outlier workers. This finding implies that only the workers who favor gamification show more engagement with the task. Another study that tested the effect of gamification on task retention and quality of the results found that retention and output quality increased when the task was gamified using levels [22]. The study tested multiple furtherance incentives and showed that game elements (badges, a leaderboard, levels, access, power, and a monetary bonus) can increase accuracy and cause tasks to be perceived as more rewarding and engaging, particularly for social incentives that involve visibility among crowd workers.

Prior work has explored the use of competitive game designs ranging from monetary reward schemes that are inspired by the success of competitions, lotteries, and games of luck to improve the cost-effectiveness of crowdsourcing tasks [76]. Rokicki et al. [77] proposed strategies for team-based crowdsourcing to improve crowdsourcing competitions, leading to performance boosts. Kobren et al. [46] proposed a survival model to predict the probability that workers will proceed to the next task available and leveraged this model to dynamically decide what task to assign and what motivating goals to present to the user. They proposed to jointly optimize for the short term (getting complex tasks done) and for the long term (keeping users engaged for more extended periods). Similarly, Gadiraju and Dietze [26] proposed using achievement priming to engage workers in long task batches and provide them with learning opportunities that can positively impact their performance. More recently, researchers proposed the use of conversational crowdsourcing as a more engaging interface for completing crowdsourcing microtasks [72] and found that using worker avatars can reduce the cognitive workload among workers and increase worker retention [68]. Inspired by prior work in this realm, we propose to leverage customizable and evolving worker avatars with features that become available to workers as they progress through task batches (giving rise to potentially evolving worker avatars).

2.2 Fostering a Sense of Community

2.2.1 Importance of Group Identification. Community identification is one of the main intrinsic motivations for crowd workers [42].

In addition, a lack of intrinsic motivation is one of the reasons why crowd workers quit their work, as intrinsic motivation starts to outweigh extrinsic motivation (often a monetary incentive) after some time [83]. Kaufmann et al. [42] describes community-based motivation as *"the acting of workers guided by the platform community, which is caused by a personal identification process"*. Furthermore, they mention social contact as another type of community-based motivation: *"motivation caused by the sheer existence of the community that offers the possibility to foster social contact"*. Their study found that the main motivators of crowd workers who spend more than 8 hours per week on MTurk are skill variety (tasks that require multiple skills that fit with the specific skill set of the worker), human capital advancement (the possibility to train useful skills), and community identification. The study of Ihl et al. [39] investigated social support (affective and instrumental), group identification, engagement, and experienced meaningfulness on crowdsourcing platforms by conducting surveys among crowd workers. Group identification was measured by the group identification scale of Doosje et al. [14]. Affective social support was measured with a questionnaire about how supported the worker felt by other crowd workers (e.g., *"The members of the crowd communities care about me."*). Instrumental social support was more focused on useful support from other workers (e.g., *"The members of the crowd communities give useful advice on job problems"*). Their main results showed that social support fosters a sense of group identification and experienced meaningfulness, contributing positively to crowd workers' subjective engagement. Corresponding to these results, through qualitative interviews with crowd workers, Soliman et al. [83] revealed that community identification is positively related to continuous participation. Thus, a sense of group identification with peers has been found to be an essential asset for motivation [42] and engagement [39, 83] in online crowd work.

2.2.2 External forums. The online solitary nature of individual crowd work tasks makes it difficult for workers to connect to their peers and foster a sense of group identification. Therefore, crowd workers often connect with peers through external platforms [31, 91, 95, 96]. These external forums help crowd workers to identify with others who do similar work, forming online communities [53]. Online communities are important as they facilitate a shared working experience among the crowd workers [85]. While these online communities serve a social goal, many crowd workers mention that their main motivation to engage in online forums is to gain information about how to optimize the quality of their work, which can optimize their earnings [58, 85, 91, 96]. Moreover, the time that crowd workers spend on these forums to gain information to improve their crowdsourcing skills is part of the 'invisible' work of crowd workers [58]. The 'invisible labor' of crowd workers refers to their work outside the tasks they perform, which is typically unpaid and unaccounted for by platforms or task requesters [30, 88]. Thus, not all workers are able or want to spend the time and effort to engage in these forums. Moreover, a study by Yin et al. [96] found that 59.1% out of 10,000 workers on MTurk reported using at least one forum, while the other 40.9% reported not being engaged on forums. Such workers cannot benefit from the social and learning opportunities that external forums offer as a community space. Therefore, researchers have suggested that social interactions between the crowd workers should be facilitated and integrated into the crowdsourcing platform itself [85, 96].

2.2.3 Fostering group identification internally. Kobayashi et al. [45] used a communication platform and a worker ranking based on the number of completed tasks to foster a sense of community. They found that fostering a sense of community positively relates to continued participation. Using such a platform increases worker visibility, which is considered to induce a sense of community and group identification [8].

We build on such prior works by attempting to foster a sense of community and increase group identification among workers completing task batches individually. To this end, we create a worker community space where workers can share their avatars and task-related feelings on successful task completion. A key difference in our effort is our focus on a lightweight intervention that does not require extensive social engagement, communication, or exchange of additional information among workers (since not all workers can indulge in such interactions and time-consuming methods can affect workers' earnings). A personalized worker avatar contributes to the ability for workers to express themselves and form their worker identity within the group of other crowd workers. We explore whether creating such visibility among crowd workers can induce a reflection on unity, causing the workers to relate to each other, thereby developing a sense of belonging [23, 84].

2.3 Hypotheses

Customizable worker avatars and avatar character selection have been shown to reduce perceived workload in information-finding tasks in conversational microtask crowdsourcing compared to conventional web interfaces without customizable worker avatars [68]. The evolution of the customizable avatars introduces a gaming element that unlocks new editable features of the avatar when the worker completes more tasks (the worker unlocks new levels). The study of Lee et al. [52] used a similar gamification approach using levels that unlock new features within a crowdsourcing task that requires workers to label cultural heritage design elements. They found that the usage of gamification reduced the perceived workload of the workers. Therefore, we expect the perceived workload to reduce when using evolving and customizable worker avatars.

> **H1a**: Evolving and customizable worker avatars will reduce the perceived workload among workers.

While the study of Qiu et al. [68] did not find any significant effects on intrinsic motivation, another study by Birk et al. [5] did find increased intrinsic motivation due to customizable avatar identification. Moreover, gamification in crowdsourcing tasks often increases motivation [62]. Specifically, prior work found that using levels in crowdsourcing tasks can improve intrinsic motivation [52]. Therefore, we expect that combining avatar customization with gamification (evolving customizable avatars) can increase intrinsic motivation. Adding gamification elements to crowdsourcing tasks can improve worker engagement [22]. We expect that increased intrinsic motivation can lead to improved subjective worker engagement. Thus, we formulate the following hypotheses:

H1b: Evolving and customizable worker avatars will lead to an increased level of intrinsic motivation.
H1c: Evolving and customizable worker avatars will lead to improved subjective worker engagement.

Prior work showed that customizable worker avatars have a positive effect on worker retention [68]. In addition, prior studies show that the willingness to complete more tasks increases as a result of gamification [21, 22, 52, 62]. Interestingly, the results presented by Maddalena et al. [56] suggest this is only the case for workers who favor gamification. Overall, we expect evolving and customizable worker avatars to increase task retention.

H2a: Evolving and customizable worker avatars will lead to increased task retention.

Prior work showed no significant improvement in task accuracy due to worker avatars [68]. Furthermore, while some prior studies suggest that data quality can be improved by gamification [22, 62], other studies did not find an increased data quality [52, 56]. The task execution time might be longer when using evolving avatars as workers might spend more time interacting with the avatar editor throughout the task.

H2b: Evolving and customizable worker avatars do not affect task accuracy.
H2c: Evolving and customizable worker avatars will lead to a longer task execution time.

Since worker avatars are known to facilitate identification [40, 68, 89], we expect that workers will identify with their avatars. Sharing and presenting their avatars in a community space with other workers can help them identify themselves as being a part of a group of crowd workers without necessarily revealing other private information. In other words, the visibility of the worker avatars might facilitate a reflection of unity [84]. Similar findings have been seen in studies about group identification and avatar customization in (serious) games [23, 86, 90]. Furthermore, the option to share their feelings about a task all workers in the cohort completed can contribute to feeling a connection with other workers [84]. Exposure to similar opinions from others has been shown to induce group identification [64].

H3: Sharing worker avatars and feelings about the task in a community space will facilitate a sense of group identification among crowd workers.

By facilitating group identification crowd workers can reflect on the fact that others are completing the same tasks as them. This notion of being part of a group might contribute to an increased intrinsic motivation of workers, which in turn can reduce their perceived workload [50]. Prior studies have found a positive relation between group or community identification and intrinsic motivation [42, 83]. As feeling part of a group can be an intrinsic motivator for workers, we expect that facilitating a community space where workers can share their avatars and task-related feelings can induce group identification and increase intrinsic motivation. Prior research found that group identification among crowd workers is positively related to user engagement [39]. Moreover, organizational identification of employees is positively related to work engagement [41]. Therefore, we expect that user engagement will be positively impacted by inducing group identification.

H4a: Creating a sense of group identification by facilitating the sharing of worker avatars and feelings reduces the perceived workload among workers.
H4b: Creating a sense of group identification by facilitating the sharing of worker avatars and feelings will lead to increased intrinsic motivation.
H4c: Creating a sense of group identification by facilitating the sharing of worker avatars and feelings improves subjective worker engagement.

As a result of the improved intrinsic motivation Kaufmann et al. [42], Kyndt et al. [50] and worker engagement Ihl et al. [39], Karanika-Murray et al. [41], we expect that facilitating a sense of group identification can increase task retention. Prior work found that community identification aids the continued participation of workers in crowd work [45]. Based on prior work, the potential of sharing worker avatars and feelings may not affect their accuracy. On the other hand, the total task execution time might be longer due to the time spent by workers in the community space.

H5a: Creating a sense of group identification by facilitating the sharing of worker avatars and feelings will lead to increased task retention.
H5b: Creating a sense of group identification by facilitating the sharing of worker avatars and feelings does not affect task accuracy.
H5c: Creating a sense of group identification by facilitating the sharing of worker avatars and feelings will lead to a longer task execution time.

3 STUDY DESIGN

To address the aforementioned research questions (**RQ1, RQ2, RQ3**), we conducted a preregistered between-subjects study with five different experimental conditions, considering two different types of tasks. In this section, we describe our overall study design, including our experimental setup, measures, and procedure in detail. Details about our technical implementation and statistical methods can be found in the Appendix, Section A.1 and A.4 respectively.

3.1 Task Design

Prior work has revealed the impact of task types on worker performance and experience-related outcomes [2, 25, 33]. To account for task type effects and better understand the generalizability of our findings, we consider two different types of tasks, an information finding task and a credibility analysis task. These types of tasks have been shown to be popular in microtask marketplaces and are commonly considered in similar studies [13, 27, 68]. Inspired

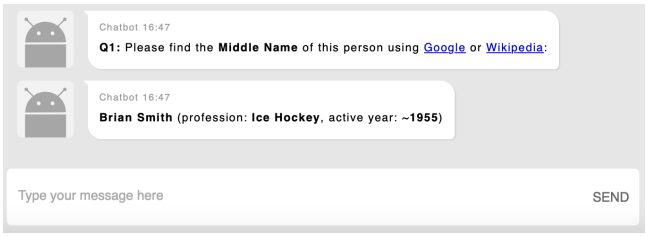


Figure 2: An example of a question from the information finding task.

by prior work that has shown that conversational crowdsourcing is an effective way to increase user engagement and satisfaction [61, 68, 72], we presented tasks to workers using a conversational interface. In both tasks, workers can refer to search on Google[5] or Wikipedia[6] to answer the question. Workers must complete at least five mandatory tasks, after which they are free to stop whenever they wish.

Information Finding: In this task, workers are asked to find the middle names of famous people by searching the Web. We used a subset of 40 questions from the dataset of Qiu et al. [69], comprising questions that provide the first and last name of a famous person, together with the profession and the active year. The task is considered to be difficult, as the dataset consists of famous people whose names and professions are similar to other famous people. To find the correct middle name, workers had to actively search based on the active year that tells these famous people apart. An example of a question can be found in Figure 2.

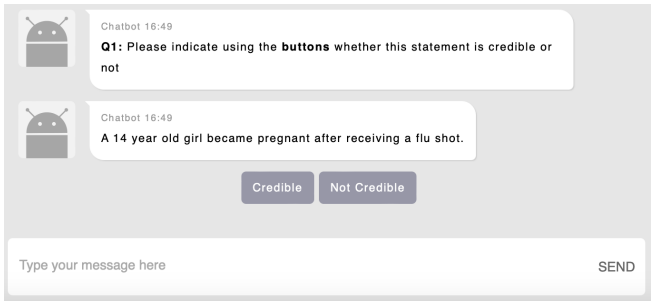
Credibility Analysis: In this task, workers are asked to read the text of statements posted online and assess their credibility — 'CREDIBLE' or 'NOT CREDIBLE.' To this end, we used the dataset compiled by Robbemond et al. [74]. The dataset consisted of 40 statements that were labeled as *credible*, *somewhat credible*, *not credible*, or *somewhat not credible*. Each category consisted of 10 statements. To increase difficulty, we combined the *somewhat credible* and the *credible* category and we combined the *somewhat not credible* and the *not credible* category. See Figure 3a for a not credible statement, and Figure 3b for a credible statement that is considered to be more difficult. The statements were ordered alphabetically to randomize the order of credibility. This resulted in a final set of 20 credible statements and 20 not credible statements.

3.2 Experimental Conditions

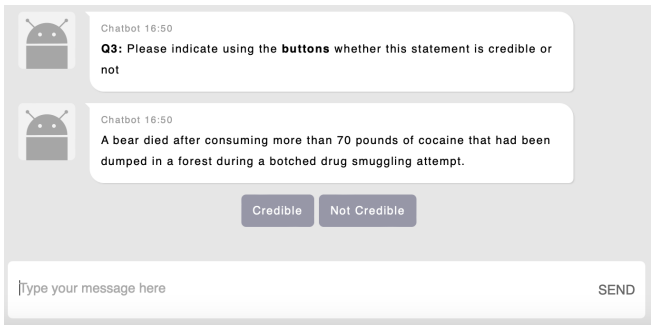
To test our hypotheses and address the research questions, we designed the following experimental conditions:

(1) *No relatable avatar* (**`Control`**): This control condition has a standard, non-human, default avatar. We expect no form of identification with this avatar. See Figure 4a for the conversational interface of this condition.
(2) *Basic avatar* (**`Basic`**): In this condition, workers are prompted with an opportunity to edit their avatar using the avatar editor before they can proceed to the tasks (cf. Section 3.2.1 and Figure 5). Workers are only able to customize basic avatar features in the avatar editor. The specifics of the avatar editor are further explained in Section 3.2.1. After starting the task, no changes can be made to the customized avatar. Workers can see their personalized avatar when working on the task in the conversational interface (see Figure 4b).
(3) *Basic avatar with community space* (**`Basic`**⊕**`Comm`**): This condition is similar to the **`Basic`** experimental condition. However, before starting the task, workers are informed that their final avatars will be shared with other crowd workers in the *worker community space* on task completion. To this end, we created a worker community space supporting different interactions (cf. Section 3.2.2 and Figure 1).
(4) *Evolving avatar* (**`Evolving`**): This condition starts similar to the **`Basic`** condition. However, for every 4 tasks, the worker unlocks a new level that reveals new editable features to further personalize the avatar. This way, we further introduce the gamification aspect to the avatar customization. Whenever a new level is unlocked, a pop-up notification shows up that notifies the worker that they have reached a new level and which features are unlocked. The worker is able to move back and forth from the avatar editor to the task to immediately check the new unlocked features. See Section 3.2.1 for a more detailed description of the avatar editor.

[5]http://www.google.com
[6]https://www.wikipedia.org

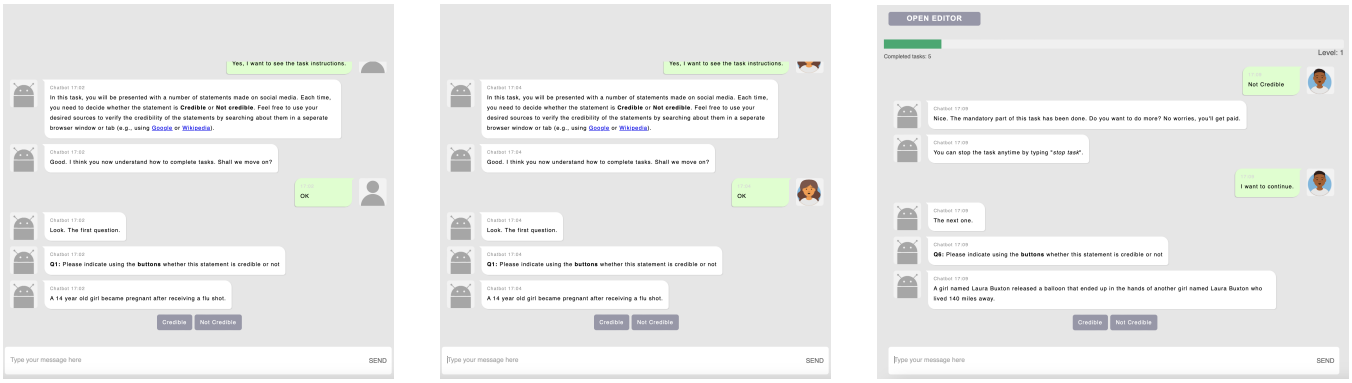


(a) Not credible



(b) Credible

Figure 3: Examples of the credibility statement questions in the credibility analysis task. Figure a) shows a statement that is not credible. Figure b) shows a credible statement that used to be a somewhat credible statement. Therefore the statement in Figure b) is considered to be difficult.



(a) `Control` **(b) `Basic` and `Basic`⊕`Comm`** **(c) `Evolving` and `Evolving`⊕`Comm`**

Figure 4: Conversational interfaces for different conditions.

(5) *Evolving avatar with community space* (**`Evolving`⊕`Comm`**): This condition is similar to the **`Evolving`** experimental condition. However, workers are informed at the beginning that on finishing their tasks, their avatars will be shared on a page with all other workers' avatars. They are informed that they can express their feelings about the task using the facial gestures of their avatar and stating how the task made them feel. By creating a space to provide visibility and expression, we aim to create a sense of group identification among the workers working on the task.

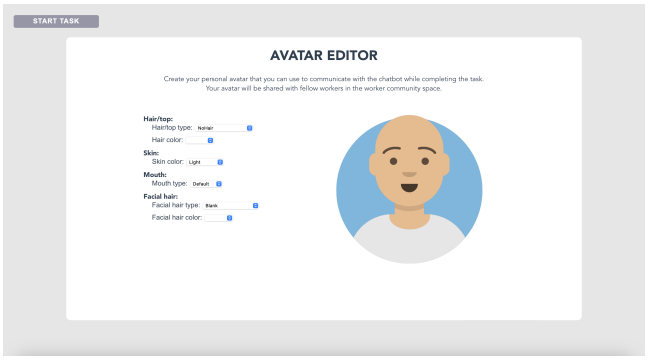


Figure 5: A screenshot showing the avatar editor interface.

3.2.1 Avatar Editor. The avatar editor is used by workers to customize their avatar prior to the task (**`Basic`**, **`Basic`⊕`Comm`**, **`Evolving`**, and **`Evolving`⊕`Comm`**), during the task (**`Evolving`** and **`Evolving`⊕`Comm`**), and after the task (**`Basic`⊕`Comm`**, **`Evolving`**, and **`Evolving`⊕`Comm`**). At the start, the avatar editor sets the avatar's eye type, mouth type, and eyebrow type to *default*. Furthermore, the initial hair/top type is set to *no hair*, and the skin color is randomly chosen. An example of the initial phase of the avatar editor can be seen in Figure 5. For the conditions including the worker community space (**`Basic`⊕`Comm`** and **`Evolving`⊕`Comm`**), an extra line of text is added to the avatar editor to notify and remind workers that their avatar will be shared with other workers on the worker community space. An overview of the initial editable features (**`Basic`** and **`Basic`⊕`Comm`**) and those that can be unlocked with new levels (**`Evolving`** and **`Evolving`⊕`Comm`**), can be found in Table 2 in the appendix along with further details of the technical implementation.

3.2.2 Worker Community Space. Workers in the community conditions (**`Basic`⊕`Comm`** and **`Evolving`⊕`Comm`**) get the opportunity to share their customized avatars and feelings about the task in the worker community space. Before entering the community space upon successful task completion, workers are given a final chance to edit and update their avatars. Workers are asked to complete a sentence with the prompt '*I am feeling ...*' by choosing a mood from the Pick-A-Mood (PAM) scale [11] (see Figure 13 in the Appendix A.3), which is displayed alongside their avatar on the community space (as shown in Figure 1). PAM is a character-based pictorial scale for reporting moods, and it has been shown to be particularly useful in capturing moods in a crowdsourcing context [70, 93, 97]. In addition, workers have the agency to choose from a variety of facial expressions to share their feelings. The moods from which workers were able to choose pertain to **pleasant** (*i.e.,* one of *excited, cheerful, relaxed, calm*), **unpleasant** (*i.e.,* one of *tense, irritated, bored, sad*), and the *neutral* mood.

We created the worker community space with the aim of fostering group identification. In the worker community space, workers see a random subset of 8 other workers' avatars and how they felt about the task. Their own avatar is placed in the middle to induce a sense of being part of the group of avatars displayed on the screen. We have implemented several interactive elements in the worker community space. Workers can use a `REFRESH` button to change the displayed subset of worker avatars at random, and the `LOAD MORE` button to display all other workers. To further increase a sense of group identification, a `SIMILAR MOOD` button was created to filter avatars of workers who reported a similar feeling. Workers in the **`Evolving`⊕`Comm`** condition were also able to order avatars based on their evolution using the `ORDER ON LEVEL` button. The worker community space only shows avatars of workers who were in the same condition and successfully completed their tasks. Furthermore, to prevent a cold start problem with a blank community space, we

added two avatars for each mood per condition to the community space (this resulted in a start with 18 avatars per condition). This design choice was made to ensure that workers could always see at least a few other avatars in the community space even when filtering on mood, with an aim to positively impact the sense of group identification among workers.

3.3 Measures

We used previously validated questionnaires to measure worker experience (*i.e.,* their perceived workload, intrinsic motivation, and subjective engagement) and group identification. When applicable, the questions were slightly altered to fit the context of our task (e.g., '*I think I did pretty well at this activity, compared to other students.*' was changed to '*I think I did pretty well at this task, compared to other workers*'). Furthermore, we measured worker retention, accuracy, and total task execution time as the task-related outcomes.

Perceived Workload. To measure the workers' perceived workload, we used the NASA-TLX [35] with a 7-point Likert scale. This questionnaire assesses workload on six different single-question dimensions. The dimensions of mental demand and physical demand describe how mentally or physically demanding the task was. Temporal demand describes how hurried or rushed the pace of the task was. The performance dimension describes how successful the worker was in accomplishing the task and the effort dimension describes how hard the worker had to work to accomplish this task. Lastly, the frustration dimension describes how insecure, discouraged, irritated, stressed, and/or annoyed the worker was when doing the task. To study the effect of evolving avatars and fostering group identification among crowd workers, we assessed the average of all dimensions (performance reversed) and each dimension separately. A high average score on the perceived workload implies that the workers perceived a high task workload.

Intrinsic Motivation. To measure the intrinsic motivation of the workers, we used three dimensions of the Intrinsic Motivation Inventory (IMI): Interest/Enjoyment (INT-ENJ), Perceived Competence (PER-COMP), and Effort/Importance (EFF-IMP) [59]. The questions were asked with a 7-point Likert scale, ranging from *1: Not at all true* to *7: Very true*. The interest/enjoyment sub-scale is considered to measure intrinsic motivation directly and consists of seven questions. The perceived competence sub-scale describes the subjective performance of the worker based on the worker's own judgment (six questions). Lastly, the sub-scale effort/importance contains five questions that address how much energy and effort the worker put into the task. Similar to the perceived workload, we analyze the average of each sub-scale separately and the total score over all sub-scales. A high score for the average overall score means that the worker has a strong intrinsic motivation to work on the task.

Subjective Engagement. To measure subjective engagement, we used the short form of the User Engagement Scale (UES-SF) with a 5-point Likert scale [65]. This scale consists of multiple subdimensions with three questions each: Focused Attention (FA; *how focused was the worker on performing the task?*), Perceived Usability (PU; *how difficult was it to interact with the task?*), Aesthetic Appeal (AE; *how attractive is the interface?*), and Reward (RW; *how rewarding was the task?*). The average score for each subdimension and the total average score are used for our analysis. A high overall subjective engagement score means that the worker was highly engaged in the task.

Group Identification. To measure the extent to which workers identify themselves as crowd workers, we used the Group Identification Measure [14]. The group identification measure consists of four questions with a 7-point Likert scale (*1: Not at all* to *7: Extremely*). The questions cover the cognitive, evaluative, and affective aspects of identification. The mean score over all four questions was measured. A high score implies a strong group identification.

In addition, to gain further insights into whether and why workers feel connected to other workers, we used a 7-point Likert scale question asking workers: *'To what extent do you feel connected to the other crowd workers that participated in this study?'*, followed by an open-ended question asking why they did or did not feel connected to the other workers. These two questions were used to code the open-ended questions into categories by two coders.

Worker Retention. To measure the objective engagement of workers in the task, we used worker retention. Worker retention is measured as the number of completed questions within one task batch. For instance, worker retention of 30 for the credibility task means that a worker classified 5 mandatory and 25 optional statements for *credible* or *not credible*. Note that there are 5 mandatory tasks and 35 additional tasks that are available within the task batch in each of the task types (*i.e.*, information finding and credibility analysis).

Worker Accuracy. For both tasks, worker accuracy is calculated as the percentage of tasks correctly completed. For the information finding task, a task is correctly completed if a worker's response contains the middle name of the famous person. For the credibility analysis task, a worker's response is considered to be correct if the right button (*i.e.*, `Credible` or `Not credible`) is pressed. Workers have the option to edit their responses to each task before their final task submission.

Task Execution Time. The task execution time is based on the total time that workers spend within the task interface (including the avatar editor, worker community space, and conversational interface). So, this is either taken from the moment the worker starts the task in the conversational interface (**`Control`**), or when the worker enters the avatar editor (all remaining conditions), up to when the worker is redirected to the post-task questionnaires.

3.4 Participant Recruitment and Procedure

Workers in our study were recruited from the Prolific crowdsourcing platform.[7] Our study was approved by the '*Human Research Ethics Committee*' of Delft University of Technology. Participation was restricted to workers who have adequate English proficiency to ensure that all workers understand the task and the questionnaires. Furthermore, workers need to be at least 18 years old. To ensure the quality of the data, we only allowed workers with an approval rate of at least 95% to participate. Workers were only allowed to participate once in our study. Based on a G-power analysis [20], the required sample size was found to be 610 workers, *i.e.,*

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

305 workers per task type; one-way ANOVA, $f = 0.2$, $\alpha = 0.05$, $power(1-\beta) = 0.8$. To account for potential exclusion due to data quality we increase the number by $\sim$ 10% to a total sample size of 680. Therefore, we recruited 340 workers per task, and 68 workers per condition within each task. Workers were paid a fair hourly wage of 9 GBP, which is above the minimum hourly wage suggested by the Prolific platform and rated as a '*good*' hourly rate on the dashboard.

Procedure. On beginning the task, workers from Prolific are redirected to a Qualtrics survey containing the informed consent. After signing the informed consent, the workers are randomly assigned to a condition and task. Subsequently, workers are redirected to the task hosted on a server. After finishing the task, the workers are directed to the post-task Qualtrics survey. Here, workers complete a set of questionnaires (cf. Section 3.3) before being redirected to Prolific on successful completion.

4 RESULTS AND ANALYSIS

4.1 Demographic Distribution

A total of 680 workers participated in our experiment, equally divided across both task types. One worker was excluded due to technical problems, and three workers were excluded due to invalid answers (all workers from the information finding task). This resulted in a final number of workers of 676 (mean age = 33.83, $SD = 11.23$). Of those workers, 61.5% identified as *male* (416 workers), 37.3% as *female* (252 workers), 1% as *non-binary* (7 workers), and 0.1% as *other* (1 worker). For the information finding task, 66 workers participated in the **`Control`** condition, 67 in the **`Basic`** condition, 68 in the **`Basic`**⊕**`Comm`** condition, 67 in the **`Evolving`** condition, and 68 in the **`Evolving`**⊕**`Comm`** condition. For the credibility analysis task, this was 68, 67, 68, 69, and 68 respectively. Descriptive statistics related to the use of the avatar editor can be found in the Appendix, Section B.1. Based on the Shapiro-Wilk tests for normality, none of our dependent measurements were normally distributed for each condition ($p < .05$). Therefore, we employed Kruskal Wallis tests to verify our hypotheses.

4.2 Perceived Workload

A non-parametric Kruskal-Wallis test was performed to investigate whether the overall TLX score and its different dimensions differ significantly across the conditions. For both tasks, the overall TLX score and the TLX dimensions did not differ across the different conditions ($\alpha = 0.05$). Thus, no significant effect was found of evolving avatars and the worker community space on workers' perceived workload.

> ***Summary: H1a)*** *We did not find any evidence for a reduced perceived workload as an effect of evolving and customizable worker avatars.* ***H4a)*** *We did not find any effect of the worker community space on workers' perceived workload. Therefore, we reject both hypotheses.*

4.3 Intrinsic Motivation

A non-parametric Kruskal-Wallis test was performed to investigate whether the overall IMI score and its dimensions differ significantly across the conditions. For both tasks, there were no significant differences found between the conditions for the overall IMI score and its subdimensions ($\alpha = 0.05$). Thus, no significant effect was found of evolving avatars and a worker community space on workers' intrinsic motivation.

> ***Summary: H1b)*** *We found no evidence of an increased intrinsic motivation as an effect of evolving and customizable worker avatars.* ***H4b)*** *Our results found no effect of a worker community space on workers' intrinsic motivation. Therefore, we reject both hypotheses.*

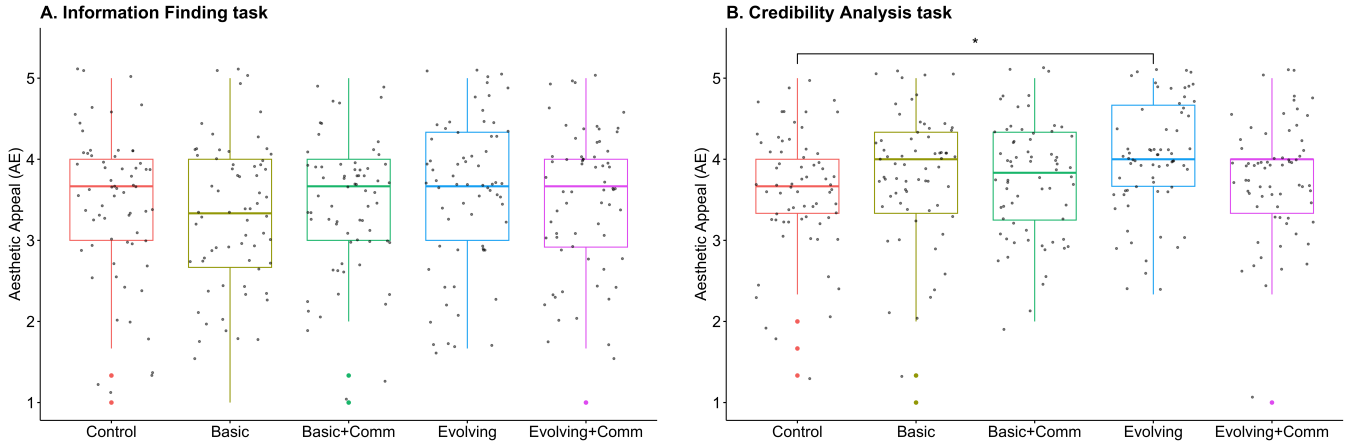
4.4 Subjective Worker Engagement

A non-parametric Kruskal-Wallis test was performed to investigate whether the overall UES score and its dimensions differ significantly across the experimental conditions (**H1c** and **H4c**). For the credibility task, we found a significant difference between conditions for the aesthetic appeal (AE) dimension ($df = 4, H = 9.739, p = .045, \alpha = 0.05$). A Dunn test was performed with a Bonferroni correction for the p-value to test which conditions differ significantly. Workers in the credibility analysis task with evolving avatars had a significantly higher aesthetic appeal score compared to workers without an avatar ($Z = -3.029, p = .025, \alpha = 0.05$; cf. Figure 6b). In contrast, there was no significant difference in aesthetic appeal for the information finding task (cf. Figure 6a).

> ***Summary: H1c)*** *Despite no significant differences found for the overall subjective engagement, workers with an evolving and customizable avatar experienced significantly greater aesthetic appeal within the credibility task. For the information finding task, no significant differences were found. Therefore, we found partial support for hypothesis 1c.* ***H4c)*** *We found no evidence of an effect of a worker community space on workers' subjective engagement. Therefore, we reject hypothesis 4c.*

4.5 Worker Retention

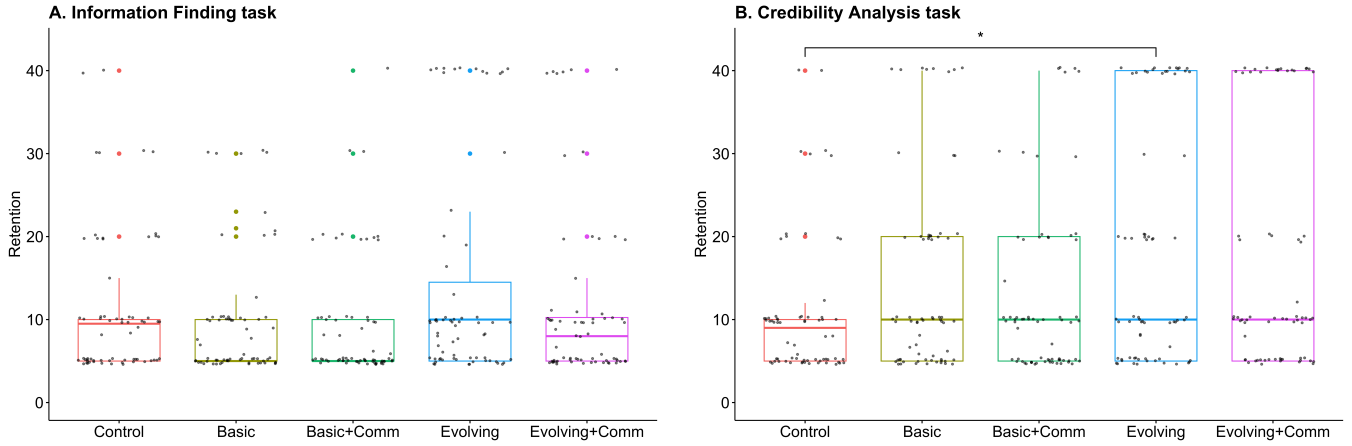
A non-parametric Kruskal Wallis test was performed to investigate whether the retention differs significantly across the conditions. The Kruskal-Wallis test showed no significant differences between the conditions for the information finding task ($H = 8.657, df = 4, p = .070, \alpha = 0.05$; see figure 7a). For the credibility analysis task, the Kruskal-Wallis test showed significant differences between the conditions ($H = 13.848, df = 4, p = .008, \alpha = 0.05$; see Figure 7b). Based on the Dunn test with a Bonferroni corrected p-value, workers with an evolving avatar had significantly higher retention than workers without an avatar ($Z = -3.121, p = .018, \alpha = 0.05$). Interestingly, workers with an evolving avatar and the worker community space did not have significantly higher worker retention compared to workers without an avatar ($Z = -2.684, p = .073$).



(a) Aesthetic Appeal (dimension of UES) for the information finding task.

(b) Aesthetic Appeal (dimension of UES) for the credibility analysis task.

Figure 6: Aesthetic Appeal (dimension of UES)



(a) Worker retention in the information finding task.

(b) Worker retention in the credibility analysis task.

Figure 7: Worker retention across the different experimental conditions and the two task types.

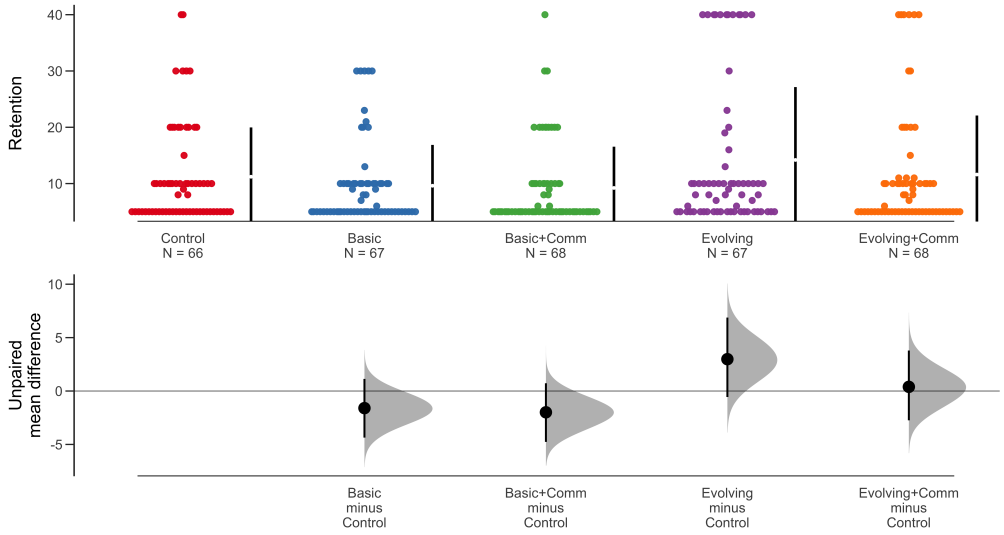
To further understand our results and their effect sizes, Figure 8 shows the estimation plots for worker retention [36]. The **`Control`** condition is compared to the other conditions. Based on these plots, we see larger effect sizes for the **`Evolving`** condition of the information finding task, and the **`Basic`**, **`Evolving`**, and **`Evolving⊕Comm`** conditions for the credibility analysis task.

Summary: H2a) *The results show that customizable and evolving worker avatars can significantly improve worker retention for the credibility analysis task. Furthermore, the estimation plots show a positive effect of evolving and customizable worker avatars across both tasks. Therefore, we found partial support for hypothesis 2a.* ***H5a)*** *We found no effect of the worker community space on worker retention. Therefore, we reject hypothesis 5a.*

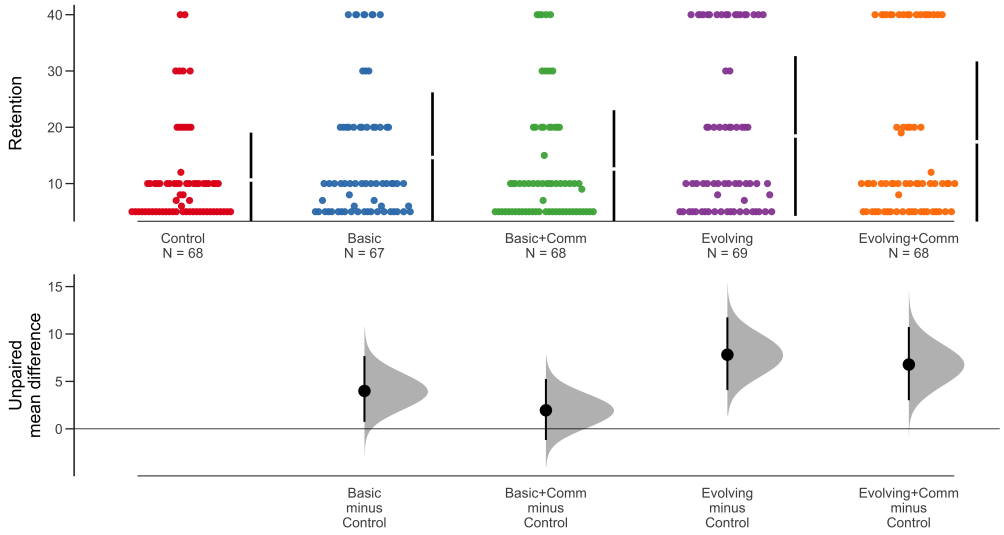
4.6 Worker Accuracy

A non-parametric Kruskal Wallis test was performed to investigate whether the accuracy differs significantly across the conditions. There were no significant differences found between the conditions for the accuracy of the information finding task ($H = 1.287, df = 4, p = 0.864$) and the credibility analysis task ($H = 4.733, df = 4, p = 0.316$).

Summary: H2b) *There is no effect found on worker accuracy as a result of evolving and customizable worker avatars.* ***H5b)*** *Likewise, the worker community space does not impact the worker's accuracy. Therefore, we accept both our hypotheses.*



(a) Information Finding task



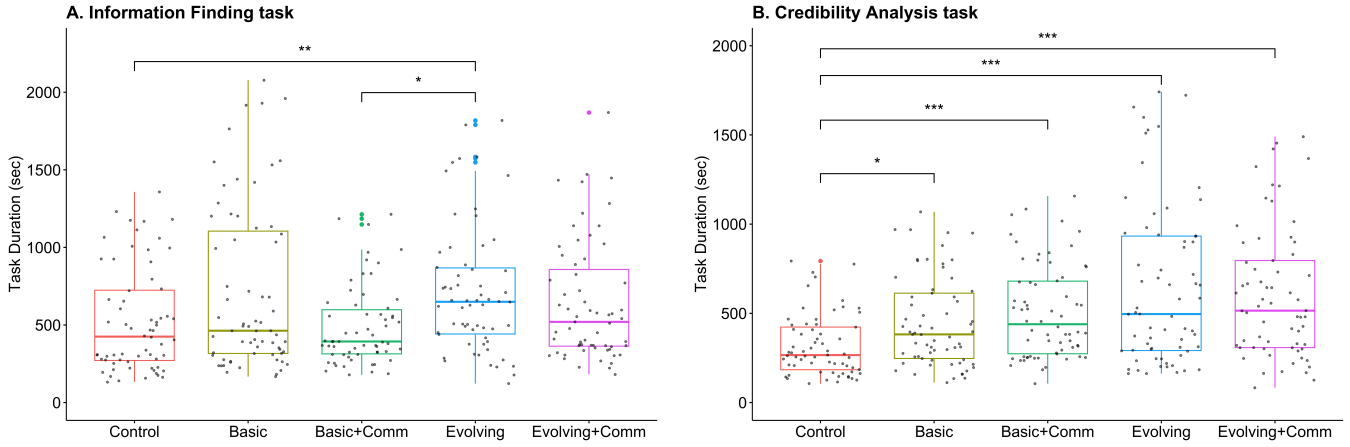
(b) Credibility Analysis task

Figure 8: Estimation plots for worker retention. For both tasks, all conditions are compared to the control condition.

4.7 Task Execution Time

For the analysis of task execution time, we removed outliers outside the whiskers of the boxplot ($Q3 + 1.5 * IQR$; $Q1 - 1.5 * IQR$) for both tasks, since these long task execution times could be an artifact of different external factors such as workers completing multiple tasks simultaneously [29], using different working strategies [33], a function of their work environments [24], and so forth. This resulted in 18 outliers being removed from the information finding task across all experimental conditions, and 12 outliers being removed from the credibility analysis task. For the information finding task, this resulted in 64 workers in the **`Control`** condition, 67 workers in **`Basic`**, 62 workers in **`Basic`**⊕**`Comm`**, 62 workers in **`Evolving`**, and 63 workers in **`Evolving`**⊕**`Comm`**. For the credibility task, this was 65, 65, 66, 67, and 65 respectively.

A Kruskal-Wallis test was performed to investigate whether there are significant differences in task duration across the conditions. The Kruskal-Wallis test revealed significant differences between the conditions for the information finding task ($H = 15.84, df = 4, p = 0.003$; cf. Figure 9a) and the credibility analysis task ($H = 36.977, df = 4, p < .001$; cf. Figure 9b). For the information finding task, the Dunn test with a Bonferroni corrected p-value showed that workers in the **`Evolving`** condition had a significantly longer task execution time than the **`Control`** condition ($Z = -3.298, p = .01, \alpha = 0.05$) and the **`Basic`**⊕**`Comm`** condition ($Z = -3.143, p = .017, \alpha = 0.05$). For the credibility analysis task, the Dunn test



(a) Task execution time for the information finding task.

(b) Task execution time for the credibility analysis task.

Figure 9: Task execution time of workers across different experimental conditions in the two task types (with outliers removed).

with a Bonferroni corrected p-value showed that workers in the **`Control`** condition had a significantly lower task execution time than workers in the **`Basic`** condition ($Z = -2.863, p = .042, \alpha = 0.05$), **`Basic`**$\oplus$**`Comm`** condition ($Z = -4.173, p < .001, \alpha = 0.05$), **`Evolving`** condition ($Z = -5.091, p < .001, \alpha = 0.05$), and the **`Evolving`**$\oplus$**`Comm`** condition ($Z = -5.207, p < .001, \alpha = 0.05$).

> ***Summary: H2c)*** *For both tasks, the task execution time is significantly longer for workers with an evolving and customizable worker avatar. Therefore, we accept hypothesis 2c.* ***H5c)*** *We found no significant effect of the worker community space on task execution time. Therefore, we reject hypothesis 5c.*

4.8 Group Identification

A non-parametric Kruskal-Wallis test was performed to investigate whether the GIM score and the connected question differ significantly across the conditions (**H3**). There were no significant differences found across conditions for the GIM score and the connected question ($\alpha = 0.05$).

To explore why workers did or did not feel connected to the other crowd workers who worked on the same tasks and whether this was related to the worker community space, the answers to the open-ended question were manually coded into categories for workers in a condition that included the worker community space. Furthermore, workers are classified based on their responses on the 7-point Likert scale as either not feeling connected ($Connected < 4$) or feeling connected ($Connected > 4$) to differentiate between the workers who felt connected or not. Open-coding was used to define different categories based on the open-ended questions of both the credibility task and the information finding task, similar to the methods of a conventional qualitative content analysis [37]. Some responses could be categorized into two different categories. The open-ended questions from both tasks were categorized using these created categories. Subsequently, a second coder used the same defined categories to categorize roughly half of the data, consisting of the open-ended questions from the credibility task ($n = 136$). A substantial inter-annotator agreement was found between the two coders, as measured with Cohen's Kappa ($\kappa = 0.744$) [51]. An overview of the description of the categories and the results can be found in the Appendix, Section B.3.

Information finding tasks. Of all the workers who worked on the information finding task that reported not feeling connected to the other workers ($n = 63$), most workers (65%, $n = 41$) did not feel connected because of a lack of direct interaction with other workers. Some workers (13%, $n = 8$) did not believe that the workers in the worker community space were indeed other workers. A smaller group of workers (6%, $n = 4$) did not feel connected because of the feelings shown in the worker community space. From the workers that did feel connected ($n = 43$), the majority of the workers felt connected because they shared a similar goal (28%, $n = 12$) or because of the feelings on the worker community space (23%, $n = 10$). A smaller fraction of the workers (9%, $n = 4$) felt connected due to the avatars in the worker community space.

Credibility analysis tasks. Of all workers from the credibility analysis task who did not feel connected to the other workers ($n = 63$), most of the workers (76%, $n = 48$) did not feel connected because there was a lack of interaction with the other workers. They felt like they were completing the tasks on their own. A smaller fraction of the workers did not feel connected because other workers mentioned they felt differently about the task (6%, $n = 4$), or the avatar was too basic an instrument to make them feel connected to other workers (6%, $n = 4$). The majority of the workers who felt connected ($n = 50$) did so because they all shared the same goal when working on the task (36%, $n = 18$). Furthermore, some workers (20%, $n = 10$) felt connected because they saw other workers reporting the same feelings about the task. Of the workers who did feel connected, a few also mentioned a lack of interaction between them and the other workers (14%, $n = 7$).

Summary: H3) *Our findings revealed that there was no significant effect of the worker community space, where workers share their avatar and feelings about the task, on either self-identification as a crowd worker or on how much they feel connected to other workers that worked on the task. Therefore, we reject our hypothesis.*

4.9 Exploratory Analysis – Group Identification

We did not find an increased sense of group identification for the conditions containing the worker community space (**H3**). With an aim to further understand group identification in our study, we explored the differences between workers who reported different levels of group identification across all conditions. To do this, we divided the workers into three groups based on their reported GIM scores: *low* ($1 \leq GIM \leq 3.5$), *mid* ($3.5 < GIM \leq 4.5$), and *high* ($4.5 < GIM \leq 7$). For the information finding task, 104 workers were found to be in the *low* group, 102 workers in the *mid* group, and 130 workers in the *high* group respectively. For the credibility analysis task, 112 workers were in the *low* group, 93 in the *mid* group, and 135 in the *high* group.

To analyze how the task duration (*i.e.,* the execution time) varied between these groups, outliers were removed from both tasks. For the information finding task, 27 outliers were removed in a similar way as described in Section 4.7, resulting in 125 workers in the *high* GIM group, 91 workers in the *mid* GIM group, and 93 workers in the *low* GIM group. For the credibility task, 18 workers were removed, resulting in 123 workers in the *high* GIM group, 91 workers in the *mid* GIM group, and 108 workers in the *low* GIM group.

4.9.1 Differences Across GIM Groups: Worker Experiences. Similar to the experimental conditions, all measurements had at least one group that did not have a normal distribution based on the Shapiro-Wilk test ($p < .05$). Therefore, we performed Kruskal-Wallis tests to investigate the differences in task-related outcomes and worker experience measurements between the different GIM groups. The results of the Kruskal-Wallis tests with all our dependent measurements can be found in Table 1. For the information finding task, we found significant differences between workers with different GIM levels for worker retention, task duration, overall TLX score (and the dimensions of mental demand, physical demand, effort, and frustration), overall IMI score (across all dimensions), and the UES score (across all dimensions). For the credibility task, we found significant differences in the accuracy, overall TLX score (the dimensions of mental demand, physical demand, and effort), overall IMI score (across all dimensions), the overall UES score (and the dimensions of FA, AE, and RW).

The results of the Dunn test for the worker experience measures, based on the Bonferroni corrected p-values, are visualized in Figure 10 (metrics for all tests can be found in the appendix, Table 6 and Table 7). For the information finding task, the workers in the *high* GIM group ($Z = 4.708, p < .001$) and the *mid* GIM group ($Z = -3.26, p = .003$) had a significantly lower TLX score than the *low* GIM group. For the credibility analysis task, the *high* GIM group had a significantly higher TLX score than the *low* GIM group ($Z = 3.64, p = .001$). For both tasks, workers in the *high* GIM

Table 1: Results for the Kruskal-Wallis test for differences across the GIM levels (*low, mid, high*). * indicates $p < .05$, ** indicates $p < .01$, and * indicates $p < 0.001$.**

Measurement	Dimension	Information Finding task		Credibility Analysis task	
		H statistic	p	H statistic	p
Retention		7.074	.029*	1.425	.49
Accuracy		2.049	.359	8.157	.017*
Task Duration		19.552	<.001***	1.239	.538
NASA-TLX		11.860	.003**	8.464	.015*
	Mental demand	15.417	<.001***	6.364	.041*
	Physical demand	13.444	.001**	10.866	.004**
	Temporal demand	4.403	.111	1.575	.455
	Performance	5.711	.058	5.389	.068
	Effort	27.575	<.001***	10.324	.006**
	Frustration	8.309	.016*	0.086	.958
IMI		82.1	<.001***	62.642	<.001***
	INT-ENJ	73.482	<.001***	46.495	<.001***
	EFF-IMP	65.735	<.001***	50.334	<.001***
	PER-COMP	32.81	<.001***	42.284	<.001***
UES		61.346	<.001***	23.168	<.001***
	FA	30.873	<.001***	8.9	.012*
	PU	12.106	.002**	1.994	.369
	AE	57.762	<.001***	47.374	<.001***
	RW	69.719	<.001***	33.09	<.001***

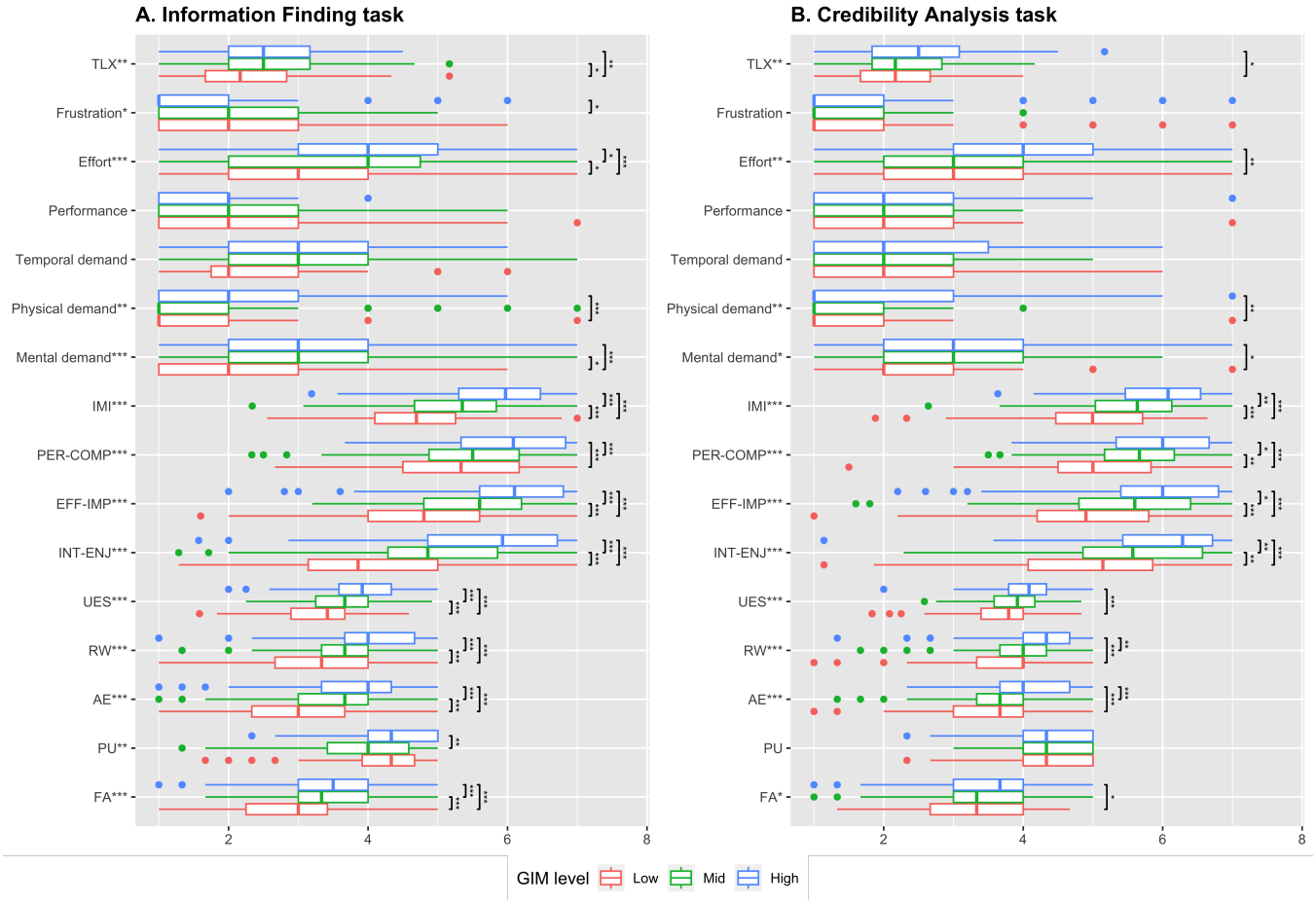


Figure 10: Worker experience measures for different levels of group identification (GIM: *low, mid, high*, represented respectively by the lower, middle, and upper boxplot per measurement). Significant differences from the Kruskal-Wallis test are shown at the measurement level (y-axis), and the significant differences (adjusted *p*-value) from the Dunn test within these measurements are shown with significance brackets between the GIM levels. * indicates $p < .05$, ** indicates $p < .01$, and * indicates $p < 0.001$. The TLX and IMI scores are measured on a 7-point Likert scale, and the UES measurements are measured on a 5-point Likert scale. Note that for the TLX measurements, a low score for the subdimension performance indicates a high perceived performance.**

group reported a significantly higher IMI score than the *mid* GIM group (information finding: $Z = 4.729, p < .001$; credibility analysis: $Z = 3.29, p = .003$) and the *low* GIM group (information finding: $Z = 9.023, p < .001$; credibility analysis: $Z = 7.914, p < .001$). Moreover, the *mid* GIM group reported significantly higher overall IMI than the *low* GIM group (information finding: $Z = -4.029, p < .001$; credibility analysis: $Z = -4.05, p < .001$). For the UES score, workers in the *high* GIM group reported significantly higher than the *low* GIM group for both tasks (information finding: $Z = 7.83, p < .001$; credibility analysis: $Z = 4.813, p < .001$). Moreover, for the information finding task, the *high* GIM group reported significantly higher than the *mid* GIM group ($Z = 3.646, p = .001$), and the *mid* GIM group reported significantly higher than the *low* GIM group ($Z = -3.931, p < .001$).

> ***Summary:*** *Workers who strongly identify themselves as a crowd worker (i.e., report high GIM scores) experience a significantly greater perceived workload but also greater intrinsic motivation and subjective engagement compared to workers who do not identify themselves with other crowd workers.*

4.9.2 Differences Across GIM Groups: Task-related Outcomes. The Dunn test with Bonferroni correction showed that workers in the *high* GIM group had significantly higher retention than workers in the *low* GIM group for the information finding task ($Z = 2.643, p = .025$; see Figure 11a). Furthermore, the task duration of the *high* GIM group was significantly longer than the task duration of the *low* GIM group ($Z = 4.162, p < .001$) and the *mid* GIM group ($Z = 3.117, p = .005$) for the information finding task (see figure 11b). For the credibility task, the accuracy of the *high* GIM group was

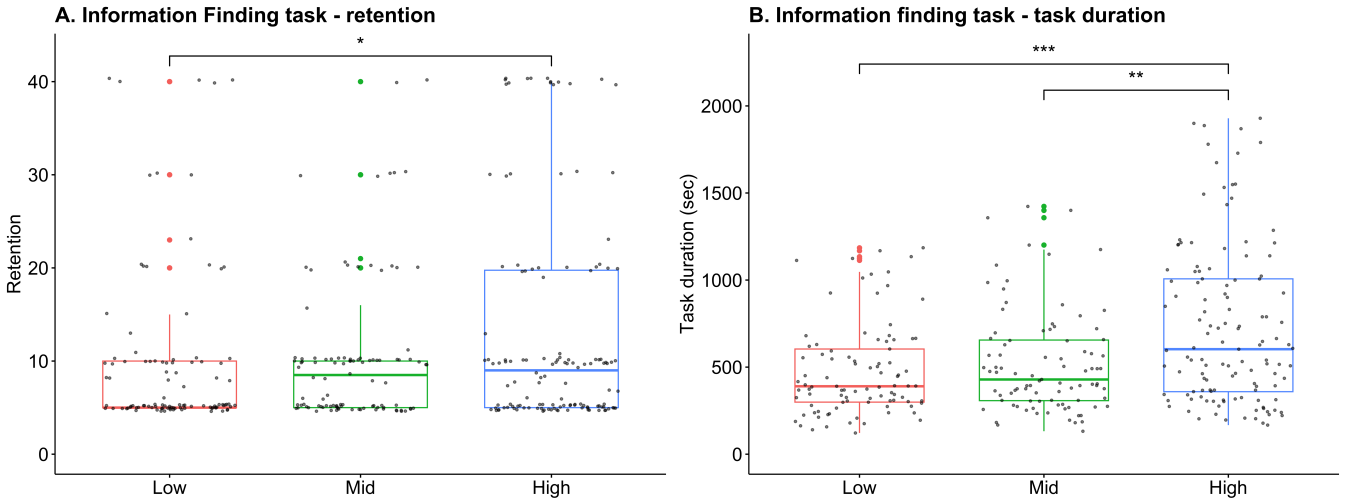


Figure 11: Significant differences between different levels of group identification (GIM: *low, mid, high*) for the task retention and duration of the information finding task. * means $p < .05$, ** means $p < .01$, and * means $p < 0.001$.**

significantly lower than the *mid* GIM group ($Z = -2.733, p = .019$; see Figure 12).

> ***Summary: i)*** *In the information finding task, workers who strongly identified as a crowd worker showed greater worker retention and task execution time.* ***ii)*** *In the credibility analysis task, workers who strongly identified as crowd worker (high group) showed less accuracy than workers who identified a little as a crowd worker (mid group).*

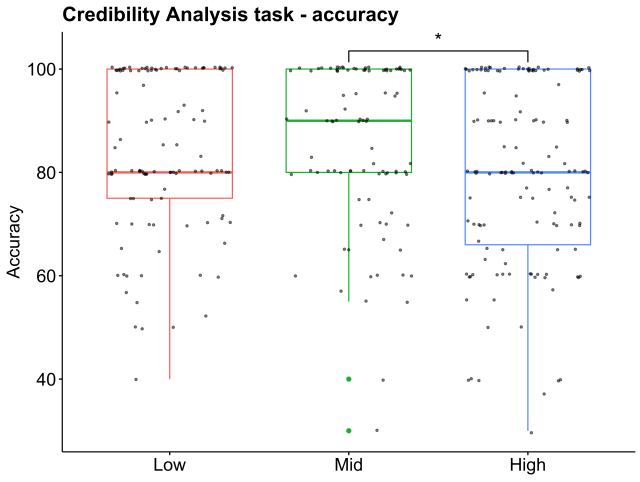


Figure 12: Significant differences between different levels of group identification (GIM: *low, mid, high*) for the accuracy of the credibility analysis task. * means $p < .05$.

4.10 Exploratory Analysis – Task Differences

Following our results which revealed differences between the credibility task and the information finding task, an exploratory analysis was carried out to further investigate how these two types of tasks were perceived differently by workers (see Figure 14 in the Appendix). Based on Wilcoxon rank tests, we found that the credibility analysis task had a significantly lower ($p = .018$) perceived workload compared to the information finding task, caused by a lower level of frustration ($p < .001$) and temporal demand ($p < .001$). Furthermore, the credibility analysis task scored higher in intrinsic motivation ($p = .004$), caused by greater interest and enjoyment ($p < .001$). In line, user engagement was greater for the credibility analysis task ($p < .001$), caused by greater perceived usefulness ($p < .001$), aesthetic appeal ($p < .001$), and reward ($p < .001$).

> ***Summary:*** *Workers in the information finding task perceived a higher workload, lower intrinsic motivation, and lower subjective engagement compared to workers in the credibility analysis task.*

5 DISCUSSION

5.1 Key Findings

5.1.1 Evolving and Customizable Avatars. The aim of our first research question (**RQ1**) was to investigate the effect of evolving and customizable worker avatars on worker experience and task-related outcomes. While we did not find any significant impact on the perceived workload, intrinsic motivation, and overall subjective engagement, our results indicate that evolving and customizable worker avatars can positively impact worker retention without decreasing accuracy. This finding is in line with prior research on the effect of avatar customization in crowdsourcing [68] and gamification in crowdsourcing [21, 22, 52, 62]. As expected, the increase in worker retention, together with some extra time that workers

use in customizing their avatars, led to a significantly increased total task execution time.

Interestingly, the increased worker retention, which can be considered an objective measurement of engagement, is not accompanied by a significant increase in subjective engagement. Only one dimension of subjective engagement, aesthetic appeal (the attractiveness of the interface), was perceived as being significantly higher for workers with an evolving and customizable avatar within the credibility analysis task. This suggests a potentially orthogonal relationship between objective worker retention and subjective worker engagement.

5.1.2 Group Identification and the Worker Community Space. We aimed to investigate whether we could foster a sense of group identification among crowd workers by providing a worker community space where workers could share their personalized worker avatar and how the tasks made them feel (**RQ2**). We proposed this as a lightweight and non-intrusive method of sharing individual information and task-related impressions to promote group identification. We expected that workers would identify with their avatar [5, 63, 89] and seeing their avatar among the other worker avatars would induce group identification [23, 84]. Our results suggest that this does not induce a statistically significant sense of group identification among the crowd workers using the worker community space. As mentioned in section 4.8, the workers who did not feel connected to other workers mainly reported a lack of interaction as the main reason. Therefore, we suggest that future work incorporates direct interaction between the crowd workers in a community space, which also resonates with prior findings related to personalized avatars and group identification in online video games [23, 86, 90]. The workers who did feel connected to other workers predominantly mentioned that sharing a goal and/or seeing the feelings of other workers on the community page made them feel connected to the other workers. The latter reason corresponds to prior work about how sharing feelings can make people feel more connected [84], and exposure to similar opinions can induce group identification [64]. However, as we did not find any significant differences in group identification and connectedness between workers in the experimental conditions with and without a community space, we expect that feeling connected and identifying with other crowd workers in our study is more likely caused by existing individual differences between the workers. Our findings suggest that workers who identify themselves as crowd workers find more meaning in the worker community space.

5.1.3 Exploratory Findings on Group Identification. Our third research question (**RQ3**) aimed to answer how a sense of group identification, induced by the worker community space, can affect worker experience and task-related outcomes. Although we did not find significant differences in the level of group identification across our experimental conditions, results from our exploratory analysis suggest that workers who strongly identify as being crowd workers experience greater intrinsic motivation and subjective engagement, corroborating prior work on group identification being related to intrinsic motivation [42, 83] and subjective engagement [39, 41]. An unexpected result is a greater perceived workload for workers who strongly identify as a crowd worker, compared to workers who do not (strongly) identify as crowd workers. A potential explanation for this could be that workers who identify themselves as crowd workers consider doing the work as an essential part of their lives and draw more meaning out of their work [97]. It is likely that those who strongly identify themselves as crowd workers also rely on crowd work for their primary livelihood (or a significant portion of their livelihood). While these workers may have greater intrinsic motivation and feel more engaged to participate in crowdsourcing tasks, their perceived workload might also be higher as they are more motivated to perform well. More research is necessary to further explore how group identification among crowd workers relates to their perceived workload, intrinsic motivation, and subjective engagement, perhaps focusing on crowd workers who spend relatively more time working on crowdsourcing tasks.

Interestingly, on exploring the relationship between group identification and task-related outcomes, we found some differences between the information finding task and the credibility analysis task. In the information finding task, we found an increased worker retention and total task execution time for workers who strongly identified as crowd workers. This finding is in line with the increased level of intrinsic motivation and subjective engagement of workers who strongly identified as crowd workers and prior work on community identification and continued participation in crowdsourcing tasks [45]. However, workers in the credibility analysis task who identified strongly as crowd workers did not exhibit an increased worker retention and task duration but exhibited a decrease in accuracy compared to workers who identified slightly as crowd workers. This suggests a potential task type-related effect, which has also been demonstrated in prior research revealing the distinct impact of different task types in crowdsourcing marketplaces [25, 68, 94].

5.1.4 Exploratory Findings on Task Differences. Our results indicate differences in the impact of evolving and customizable avatars and group identification between the information finding and credibility analysis tasks. Workers in the credibility analysis task show significantly greater worker retention due to evolving and customizable worker avatars. The results of the information finding task do not show a significant effect, but the results indicate a positive effect on worker retention (see section 4.5). A similar effect is seen for the workers in the experimental conditions with evolving avatars and the worker community space. We found that workers in the credibility analysis task reported a significantly higher perception of aesthetic appeal, which was not the case for the information finding task. Our exploratory findings for the different GIM levels also revealed differences in the task-related outcomes between the two task types.

These differences in our findings across the tasks suggest that there might be an important role for task features that can either mitigate or amplify the impact of evolving avatar customization or group identification. Based on prior work, some task features that could have influenced this effect may be the task complexity, enjoyment, and/or the effort to come up with an answer [94]. We carried out an exploratory analysis to understand potential differences in worker perceptions of the credibility analysis and the information finding task. This analysis revealed that the credibility analysis task was perceived as less frustrating, less hurried/rushed, inducing greater interest and enjoyment, being less difficult to interact with,

having a more attractive interface, and being more rewarding than the information finding task. These differences may have mitigated the impact of the evolving and customizable avatars in the information finding task on worker retention and the perception of the attractiveness of the task interface. Furthermore, we saw that workers who identify strongly as crowd workers put in more work and time in a task that is generally perceived as more frustrating and less enjoyable (the information finding task). For a more enjoyable task (the credibility task), workers who did not (strongly) identify put in the same amount of work and time as those who strongly identified as crowd workers. Future research can further explore the role of task types in the effectiveness of gamification interventions and the effect of group identification.

5.2 Caveats, Limitations, and Other Considerations

Novelty Effect. It is possible that the effects we observed as a result of gamifying the avatar customization by tying it together with task progress is caused by a novelty effect, and may not be sustainable over a long-term [32]. Such novelty effects often occur for gamification that is focused on extrinsic game elements [75]. However, we chose evolving and customizable worker avatars because it is an extrinsic game element and is therefore not bound to a specific crowdsourcing task context. Future work is necessary to determine whether this approach can reap continued benefits over a long term. For instance, incorporating evolving customizable avatars in a crowdsourcing platform and/or integrating them within a permanent or dynamic worker community space can ensure that any progress made by workers does not get lost beyond the task itself. This way, the virtual worker identity formed by the avatar is maintained over time by the integration of the crowdsourcing platform itself. Perhaps future work could investigate how this virtual identity can contribute to more elaborate social interactions that can be implemented directly in crowdsourcing tasks and platforms.

Potential Biases. Cognitive biases can negatively impact the outcome of crowdsourcing experiments [18, 38, 80]. We used the Cognitive Bias Checklist to analyze and report potential biases in our study [15]. *Confirmation bias* may have surfaced in our work through the credibility analysis tasks that we considered. The statements used in the credibility analysis task could relate to a worker's prior beliefs about specific topics. For instance, a worker who identifies as an anti-vaxer might have a confirmation bias to flag the statement *'The CDC issued a warning to all Americans urging them not to get the flu shot this year.'* as being 'CREDIBLE.' Another potential cognitive bias that may have surfaced is *loss aversion*. Although we mentioned to the workers that they would get paid based on an hourly wage, workers may have chosen to drop out of the task batch earlier to ensure their earnings. Prior work in crowdsourcing literature has identified and corroborated such behavior [34]. While both cognitive biases could have influenced the task-related outcomes, it is unlikely that these biases have caused significant differences across the experimental conditions and, therefore, may not affect the validity of conclusions drawn in this study.

Ethical Issues and Considerations. Shahri et al. [81] identified different ethical issues that can be caused by deploying gamification techniques in a workplace. Some ethical issues raised are related to leaderboards, privacy, exploitation, and personal and cultural values. Within our study, the functionality within the worker community space to order workers based on the levels reached might have caused workers to feel bad about their performance and their relatively less evolved avatars. However, this effect may have been mitigated by ensuring the anonymity of workers.

Furthermore, the worker community space is limited to serve workers who completed the task successfully, which might conflict with personal and/or cultural values. It can be considered unfair towards other workers, as fostering group identification and increasing worker experience can be seen as a right for all workers. Future work could investigate ways to foster group identification during and before tasks to deal with this value conflict. From a task requesters' perspective, fostering a sense of group identification before task completion might also benefit engagement during the task [39]. Another ethical issue related to gamification in a workplace is whether increasing workers' productivity with gamification is exploitative [43, 81]. As observed in our study, gamification might cause workers to complete more work. This is a problem when workers are not paid for their extra efforts or suffer due to the workload. We argue that there is positive value in employing gamification to increase productivity, aiming to improve workers' experience and motivation to engage in the work [52, 62]. However, increasing productivity should not cause an excessive workload or affect the short and long-term health of workers [4], and workers should be paid fair wages [92].

Platform Differences. As our current study focused on the Prolific crowdsourcing platform, we are unsure how the results generalize towards other platforms, such as Amazon Mechanical Turk (MTurk), Appen, or Toloka. Different platforms have differences in how they are used, the number of hours that workers generally spend on the platforms, and their workers' demographic and geographic features [67]. Moreover, some workers are active on multiple crowdsourcing platforms. Future research can investigate the potential platform-specific needs of workers and how to facilitate an appropriate working identity that suits worker needs.

5.3 Implications and Future Work

Our work has important design and theoretical implications, which we discuss in detail in this section.

Evolving and Customizable Worker Avatars for Crowdsourcing Tasks. Our findings have important implications for the design of future crowdsourcing microtasks. Task requesters often desire worker retention in tasks with elaborate training or tutorial phases. Based on our results, evolving and customizable worker avatars in monotonous crowdsourcing tasks can improve worker retention in conversational crowdsourcing. Though the evolving aspect of the customizable worker avatars can lead to an increased focus on completing more microtasks among workers, our results suggest that accuracy is not negatively impacted. Prior crowdsourcing literature has also revealed a positive impact of increased worker retention on overall accuracy [26]. Furthermore, evolving worker avatars can be particularly interesting when designing crowdsourcing tasks where worker retention plays an important role. For instance, tasks that require training or tutorials.

In that case, increasing worker retention might save costs related to training the worker. Considering the benefits that can be reaped from worker retention in long batches of tasks (such as learning effects, improvement in accuracy, task efficiency, and stable performance), this method shows the potential to improve worker experiences while meeting task requesters' needs. Additionally, the context-independent nature of integrating evolving and customizable worker avatars makes this viable for different tasks. Our results, however, indicate that task-specific features can play a role in mediating the effect of customizable worker avatars and group identification. Future work is necessary to investigate how and the extent to which task-dependent features shape the impact of evolving and customizable avatars in fostering group identification and shaping task-related outcomes and worker experience.

Group Identification and Sustainable Crowd Work. Our exploratory findings have highlighted the importance of improving group identification among crowd workers working on individual crowdsourcing tasks. This has important theoretical and practical implications for the broad context of crowdsourcing. Our results indicate that group identification is related to greater intrinsic motivation and subjective engagement. Based on this, we believe that fostering group identification contributes positively to the worker experience, which can help create a stronger and thriving workforce [44]. Therefore, we envision that fostering group identification can aid in improving the sustainability of crowd work. While workers who identified themselves strongly as crowd workers showed greater intrinsic motivation and subjective engagement, they also experienced a greater perceived workload. These findings highlight important future directions for optimizing a healthy and sustainable work environment for crowd workers. Future work can further explore effective means to foster a sense of community among crowd workers who predominantly work on tasks individually. More work is needed to understand how we can increase workers' intrinsic motivation and engagement while maintaining a healthy level of perceived workload.

6 CONCLUSIONS

Our first research question was to investigate the effect of evolving and customizable worker avatars on worker experience and task-related outcomes (**RQ1**). To address this question, we created a conversational crowdsourcing task where workers were able to customize their worker avatars, and as they progressed through the task batches, they unlocked new levels that allowed them to use new features to customize their avatars. We measured task-related outcomes, such as worker retention, accuracy, and total task execution time. The worker experience was measured by perceived workload, intrinsic motivation, and subjective engagement. Our results suggest that evolving and customizable worker avatars can increase worker retention. Our second research question addressed the extent to which the sharing of worker avatars and task-related feelings in a worker community space could foster a sense of group identification among crowd workers (**RQ2**). We created an interactive worker community space where workers shared their personalized worker avatars with their feelings on the task. However, the worker community space did not successfully foster an increased sense of group identification among crowd workers, although exploratory findings revealed that this could be a function of individual differences among crowd workers. With our third research question, we investigated the effect of group identification, induced by the worker community space, on worker experience and task-related outcomes (**RQ3**). We found that the worker community space did not improve group identification among the crowd workers. We conducted an exploratory analysis to investigate the effect of different levels of group identification across all workers on task-related outcomes and worker experience. Our results indicated that workers who identify themselves as crowd workers experience a significantly greater perceived workload, intrinsic motivation, and subjective engagement. Our study contributes to extending the understanding of designing future crowdsourcing tasks. It sheds light on new directions to improve the sustainability of the crowdsourcing paradigm for crowd workers, task requesters, and crowdsourcing platforms.

ACKNOWLEDGMENTS

This research is (partially) funded by the Convergence, the alliance between Erasmus Medical Centre Rotterdam, Erasmus University Rotterdam, and Delft University of Technology. We thank Mohammed Al Owayyed for his contributions to the qualitative analysis in this work. We also thank all participants from Prolific and the anonymous reviewers for their valuable feedback.

REFERENCES

[1] Tahir Abbas and Ujwal Gadiraju. 2022. Goal-Setting Behavior of Workers on Crowdsourcing Platforms: An Exploratory Study on MTurk and Prolific. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 2–13.

[2] Tahir Abbas, Ujwal Gadiraju, Vassilis-Javed Khan, and Panos Markopoulos. 2021. Making Time Fly: Using Fillers to Improve Perceived Latency in Crowd-Powered Conversational Systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9. 2–14.

[3] Tahir Abbas, Vassilis-Javed Khan, Ujwal Gadiraju, and Panos Markopoulos. 2020. Trainbot: A conversational interface to train crowd workers for delivering on-demand therapy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 3–12.

[4] Garrett Allen, Andrea Hu, and Ujwal Gadiraju. 2022. Gesticulate for Health's Sake! Understanding the Use of Gestures as an Input Modality for Microtask Crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 14–26.

[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* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2982–2995. https://doi.org/10.1145/2858036.2858062

[6] Nathan A Bowling, Kevin J Eschleman, and Qiang Wang. 2010. A meta-analytic examination of the relationship between job satisfaction and subjective well-being. *Journal of Occupational and Organizational Psychology* 83, 4 (2010), 915–934.

[7] Daren C Brabham, Kurt M Ribisl, Thomas R Kirchner, and Jay M Bernhardt. 2014. Crowdsourcing applications for public health. *American journal of preventive medicine* 46, 2 (2014), 179–187.

[8] John T Bush and Rachel M Balven. 2021. Catering to the crowd: An HRM perspective on crowd worker engagement. *Human Resource Management Review* 31, 1 (2021), 100670.

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

[10] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2334–2346.

[11] 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.

[12] 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.

[13] 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.

[14] Bertjan Doosje, Naomi Ellemers, and Russell Spears. 1995. Perceived intragroup variability as a function of group status and identification. *Journal of experimental social psychology* 31, 5 (1995), 410–436.

[15] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 9. 48–59.

[16] Olive Jean Dunn. 1964. Multiple comparisons using rank sums. *Technometrics* 6, 3 (1964), 241–252.

[17] Tom Edixhoven, Sihang Qiu, Lucie Kuiper, Olivier Dikken, Gwennan Smitskamp, and Ujwal Gadiraju. 2021. Improving reactions to rejection in crowdsourcing through self-reflection. In *Proceedings of the 13th ACM Web Science Conference 2021*. 74–83.

[18] Carsten Eickhoff. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 162–170.

[19] Shaoyang Fan, Ujwal Gadiraju, Alessandro Checco, and Gianluca Demartini. 2020. Crowdco-op: Sharing risks and rewards in crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–24.

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

[21] Yuanyue Feng, Hua Jonathan Ye, Ying Yu, Congcong Yang, and Tingru Cui. 2018. Gamification artifacts and crowdsourcing participation: Examining the mediating role of intrinsic motivations. *Computers in Human Behavior* 81 (2018), 124–136.

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

[23] Alessandro Gabbiadini, Silvia Mari, Chiara Volpato, and Maria Grazia Monaci. 2014. Identification processes in online groups. *Journal of Media Psychology* (2014).

[24] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. 2017. Modus operandi of crowd workers: The invisible role of microtask work environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–29.

[25] 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 (2019), 815–841.

[26] Ujwal Gadiraju and Stefan Dietze. 2017. Improving learning through achievement priming in crowdsourced information finding microtasks. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. 105–114.

[27] 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.

[28] Ujwal Gadiraju and Jie Yang. 2020. What can crowd computing do for the next generation of AI systems?. In *2020 Crowd Science Workshop: Remoteness, Fairness, and Mechanisms as Challenges of Data Supply by Humans for Automation*. CEUR, 7–13.

[29] Sandy JJ Gould, Anna L Cox, and Duncan P Brumby. 2016. Diminished control in crowdsourcing: An investigation of crowdworker multitasking behavior. *ACM Transactions on Computer-Human Interaction (TOCHI)* 23, 3 (2016), 1–29.

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

[31] Mary L Gray, Siddharth Suri, Syed Shoaib Ali, and Deepti Kulkarni. 2016. The crowd is a collaborative network. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 134–147.

[32] Juho Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does gamification work?–a literature review of empirical studies on gamification. In *2014 47th Hawaii international conference on system sciences*. Ieee, 3025–3034.

[33] Lei Han, Eddy Maddalena, Alessandro Checco, Cristina Sarasua, Ujwal Gadiraju, Kevin Roitero, and Gianluca Demartini. 2020. Crowd worker strategies in relevance judgment tasks. In *Proceedings of the 13th international conference on web search and data mining*. 241–249.

[34] 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* 33, 5 (2019), 2266–2279.

[35] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[36] 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.

[37] Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research* 15, 9 (2005), 1277–1288.

[38] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[39] Andreas Ihl, Kim Simon Strunk, and Marina Fiedler. 2020. The mediated effects of social support in professional online communities on crowdworker engagement in micro-task crowdworking. *Computers in Human Behavior* 113 (2020), 106482.

[40] Dominic Kao and D Fox Harrell. 2018. The effects of badges and avatar identification on play and making in educational games. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–19.

[41] Maria Karanika-Murray, Nikita Duncan, Halley M Pontes, and Mark D Griffiths. 2015. Organizational identification, work engagement, and job satisfaction. *Journal of Managerial Psychology* 30, 8 (2015), 1019–1033.

[42] N. Kaufmann, Thimo Schulze, and Daniel Veit. 2011. More than fun and money. Worker Motivation in Crowdsourcing–A Study on Mechanical Turk. *Proceedings of the Seventeenth Americas Conference on Information Systems*.

[43] Tae Wan Kim. 2018. Gamification of labor and the charge of exploitation. *Journal of business ethics* 152 (2018), 27–39.

[44] 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.

[45] Masatomo Kobayashi, Shoma Arita, Toshinari Itoko, Shin Saito, and Hironobu Takagi. 2015. Motivating multi-generational crowd workers in social-purpose work. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 1813–1824.

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

[47] Jonna Koivisto and Juho Hamari. 2014. Demographic differences in perceived benefits from gamification. *Computers in Human Behavior* 35 (2014), 179–188.

[48] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123 (2017), 32–73.

[49] William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* 47, 260 (1952), 583–621.

[50] Eva Kyndt, Inneke Berghmans, Filip Dochy, and Lydwin Bulckens. 2014. 'Time is not enough.'Workload in higher education: a student perspective. *Higher Education Research & Development* 33, 4 (2014), 684–698.

[51] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[52] Jieun Lee, Ji Hyun Yi, and Seungjun Kim. 2020. Cultural heritage design element labeling system with gamification. *IEEE Access* 8 (2020), 127700–127708.

[53] Vili Lehdonvirta and Paul Mezier. 2013. Identity and self-organization in unstructured work. *Dynamics of Virtual Work Working Paper Series* 1 (2013), 1–35.

[54] Howard Levene. 1960. Robust tests for equality of variances. *Contributions to probability and statistics* (1960), 278–292.

[55] Ioanna Lykourentzou, Shannon Wang, Robert E Kraut, and Steven P Dow. 2016. Team dating: A self-organized team formation strategy for collaborative crowdsourcing. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. 1243–1249.

[56] Eddy Maddalena, Luis-Daniel Ibáñez, Neal Reeves, and Elena Simperl. 2023. Qrowdsmith: Enhancing paid microtask crowdsourcing with gamification and furtherance incentives. *ACM Transactions on Intelligent Systems and Technology* 14, 5 (2023), 1–26.

[57] Anoush Margaryan, Timothy Charlton, and Ujwal Gadiraju. 2020. Learning and skill development in online platform work: Comparing Microworkers' and Online Freelancers' Practices (CrowdLearnPlus). In *Copenhagen Business School, CBS*.

[58] 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.

[59] 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.

[60] 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.

[61] Robert J Moore, Raphael Arar, Guang-Jie Ren, and Margaret H Szymanski. 2017. Conversational UX design. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. 492–497.

[62] 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.

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

[64] Caoimhe O'Reilly, Paul J Maher, Adrian Lüders, and Michael Quayle. 2022. Sharing is caring: How sharing opinions online can connect people into groups and foster identification. *Acta Psychologica* 230 (2022), 103751.

[65] Heather L O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28–39.

[66] Zhengxiang Pan, Han Yu, Chunyan Miao, and Cyril Leung. 2016. Efficient collaborative crowdsourcing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

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

[68] Sihang Qiu, Alessandro Bozzon, Max V. Birk, and Ujwal Gadiraju. 2021. Using Worker Avatars to Improve Microtask Crowdsourcing. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 322 (oct 2021), 28 pages. https://doi.org/10.1145/3476063

[69] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Estimating conversational styles in conversational microtask crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.

[70] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Just the right mood for HIT! Analyzing the role of worker moods in conversational microtask crowdsourcing. In *Web Engineering: 20th International Conference, ICWE 2020, Helsinki, Finland, June 9–12, 2020, Proceedings 20*. Springer, 381–396.

[71] 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.

[72] Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon, and Geert-Jan Houben. 2020. Conversational Crowdsourcing.. In *CSW@ NeurIPS*. 1–6.

[73] Will Reese. 2008. Nginx: the high-performance web server and reverse proxy. *Linux Journal* 2008, 173 (2008), 2.

[74] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the Role of Explanation Modality in AI-assisted Decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 223–233.

[75] Luiz Rodrigues, Filipe D Pereira, Armando M Toda, Paula T Palomino, Marcela Pessoa, Leandro Silva Galvão Carvalho, David Fernandes, Elaine HT Oliveira, Alexandra I Cristea, and Seiji Isotani. 2022. Gamification suffers from the novelty effect but benefits from the familiarization effect: Findings from a longitudinal study. *International Journal of Educational Technology in Higher Education* 19, 1 (2022), 1–25.

[76] Markus Rokicki, Sergiu Chelaru, Sergej Zerr, and Stefan Siersdorfer. 2014. Competitive game designs for improving the cost effectiveness of crowdsourcing. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. 1469–1478.

[77] 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.

[78] J Patrick Royston. 1982. An extension of Shapiro and Wilk's W test for normality to large samples. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 31, 2 (1982), 115–124.

[79] G Rupert Jr et al. 2012. Simultaneous statistical inference. (2012).

[80] Farah Saab, Imad H Elhajj, Ayman Kayssi, and Ali Chehab. 2019. Modelling cognitive bias in crowdsourcing systems. *Cognitive Systems Research* 58 (2019), 1–18.

[81] Alimohammad Shahri, Mahmood Hosseini, Keith Phalp, Jacqui Taylor, and Raian Ali. 2014. Towards a code of ethics for gamification at enterprise. In *The Practice of Enterprise Modeling: 7th IFIP WG 8.1 Working Conference, PoEM 2014, Manchester, UK, November 12-13, 2014. Proceedings 7*. Springer, 235–245.

[82] Shahin Sharifi Noorian, Sihang Qiu, Burcu Sayin, Agathe Balayn, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2023. Perspective: Leveraging Human Understanding for Identifying and Characterizing Image Atypicality. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 650–663.

[83] Wael Soliman, Tapani Rinta-Kahila, and Joona Kaikkonen. 2019. Why Is Your Crowd Abandoning You?: Exploring Crowdsourcing Discontinuance through the Lens of Motivation Theory. In *Australasian Conference on Information Systems*. Association for Information Systems.

[84] Ekaterina R Stepanova, John Desnoyers-Stewart, Kristina Höök, and Bernhard E Riecke. 2022. Strategies for Fostering a Genuine Feeling of Connection in Technologically Mediated Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–26.

[85] Kim Simon Strunk and Franz Strich. 2023. Building professional holding environments for crowd work job crafting through online communities. *Information Systems Journal* (2023).

[86] Masanori Takano and Fumiaki Taka. 2022. Fancy avatar identification and behaviors in the virtual world: Preceding avatar customization and succeeding communication. *Computers in Human Behavior Reports* 6 (2022), 100176.

[87] Lieke L ten Brummelhuis, Claartje L Ter Hoeven, Arnold B Bakker, and Bram Peper. 2011. Breaking through the loss cycle of burnout: The role of motivation. *Journal of Occupational and Organizational Psychology* 84, 2 (2011), 268–287.

[88] Carlos Toxtli, Siddharth Suri, and Saiph Savage. 2021. Quantifying the invisible labor in crowd work. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–26.

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

[90] Jan Van Looy, Cédric Courtois, and Melanie De Vocht. 2010. Player identification in online games: Validation of a scale for measuring identification in MMORPGs. In *Proceedings of the 3rd International Conference on Fun and Games*. 126–134.

[91] Xinyi Wang, Haiyi Zhu, Yangyun Li, Yu Cui, and Joseph Konstan. 2017. A community rather than a union: Understanding self-organization phenomenon on Mturk and how it impacts Turkers and requesters. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2210–2216.

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

[93] Luyan Xu, Xuan Zhou, and Ujwal Gadiraju. 2019. Revealing the role of user moods in struggling search tasks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1249–1252.

[94] Jie Yang, Judith Redi, Gianluca Demartini, and Alessandro Bozzon. 2016. Modeling task complexity in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 4. 249–258.

[95] Jie Yang, Carlo van der Valk, Tobias Hoßfeld, Judith Redi, and Alessandro Bozzon. 2018. How do crowdworker communities and microtask markets influence each other? A data-driven study on Amazon Mechanical Turk. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 6. 193–202.

[96] Ming Yin, Mary L Gray, Siddharth Suri, and Jennifer Wortman Vaughan. 2016. The communication network within the crowd. In *Proceedings of the 25th International Conference on World Wide Web*. 1293–1303.

[97] 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.

A STUDY DESIGN

A.1 Technical Implementation

We used *TickTalkTurk* [71] to design the conversational task interface and leveraged a Vue.js library[8] of the Avataaars library[9] to create an avatar editor for workers. The front end of the task interface, including the avatar editor, conversational interface, and the worker community space was built using the JavaScript Framework Vue.js.[10] The back end was built with Flask [11] in Python and connected to a MongoDB database.[12] The application was hosted on a Ubuntu 22.04 server using Nginx [73] and Gunicorn,[13] and secured with an SSL certificate by Let's Encrypt.[14]

A.2 Editable Features of Avatars

An overview of the editable features in the avatar editor can be found in Table 2.

A.3 Pick-A-Mood Scale

Figure 13 shows the interface where workers are asked how the task made them feel based on the Pick-A-Mood scale [11] before entering the worker community space.

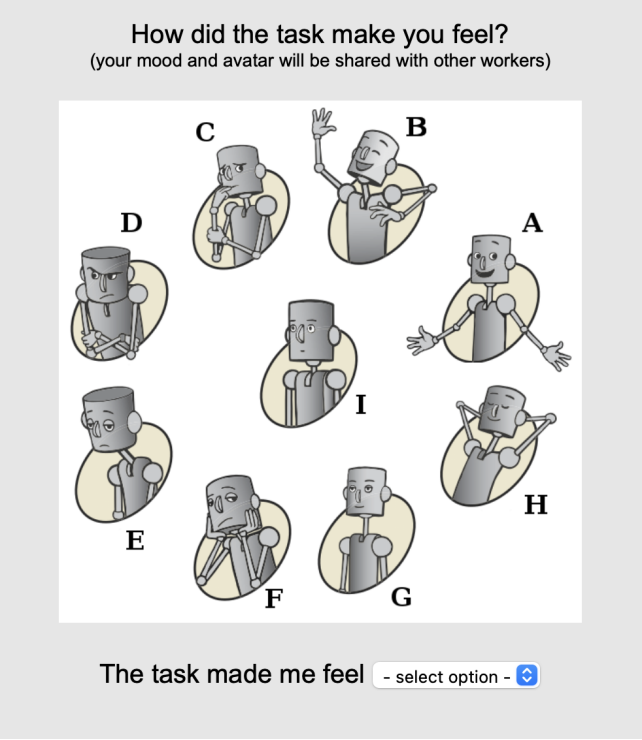


Figure 13: Workers are asked how the task made them feel, based on the Pick-A-Mood scale.

[8]https://github.com/orgordin/vuejs-avataaars
[9]https://getavataaars.com
[10]https://vuejs.org
[11]https://flask.palletsprojects.com/en/2.3.x/
[12]https://www.mongodb.com
[13]https://gunicorn.org
[14]https://letsencrypt.org

A.4 Statistical Analysis

To test our hypotheses, we want to compare the conditions for each dependent variable that is related to the crowd worker experience or task-related outcomes. For each dependent variable, we tested whether each condition is normally distributed using a Shapiro-Wilk test [78]. If the dependent variable is normally distributed across all conditions, we test the homogeneity of variances among the conditions with Levene's test [54].

If the assumption of normality and homogeneity of variances were met, a one-way ANOVA test was performed to test for significant differences between the conditions. If the assumptions are not met, a Kruskal-Wallis test was performed [49]. To further investigate the differences between the conditions, post-hoc tests were carried out, while appropriately adjusting for multiple comparisons to avoid type-I error inflation. For the parametric one-way ANOVA test, Tukey's test [79] was performed. In the case of the non-parametric Kruskal-Wallis, a Dunn test [16] was performed. As demographic differences can influence the effect of gamification [47], we explored potential confounds of age and/or gender by carrying out corresponding ANCOVA tests while considering these variables as covariates. These results can be found in Section B.2.

B RESULTS

B.1 Descriptive Statistics

Table 3 shows the descriptive statistics of the avatar editor and the worker community space to gain insights into how workers interacted with the avatar editor and the community space. As expected, the number of changes made in the avatar editor is higher for the evolving avatar conditions. Furthermore, the descriptive results show that workers actively customized their avatars. The descriptive statistics of the worker community indicate that on average, the workers did not interact much with the buttons, while they did spend some time in the worker community space.

B.2 Covariance Analysis

To verify whether gender and age played a role in shaping the significant differences we found in worker retention and aesthetic appeal across the different experimental conditions for the credibility analysis task, we performed an ANCOVA test between all conditions, using gender and age as covariates. For worker retention, our ANCOVA test does not show any effect of age ($df = 1, F = 1.357, p = .245$) or gender ($df = 3, F = 0.613, p = .607$). The ANCOVA test for aesthetic appeal does not show any significant effect of age ($df = 1, F = 0.610, p = .285$) but does show a significant effect of gender ($df = 3, F = 2.692, p = .046, \alpha = 0.05$). A post-hoc Tukey test revealed a significant difference between the categories *non-binary* and *other* ($p = .026$). These two categories only consist of 3 and 1 worker, respectively. Therefore, we can conclude that the variables of age and gender did not affect our findings.

B.3 Group Identification – Qualitative Analysis

The description of the categories that emerged from the open-coding of the responses on the open-ended question about group identification can be found in Table 4, together with an example

Table 2: All the editable features of the avatar editor. The 'Basic Items' are the items that are always available in the avatar editor. The 'Evolving Items' are the features that can be unlocked by reaching a new level. When these items are unlocked are shown in the 'Unlocked' column. These items are based on the Avataaars generator (https://getavataaars.com).

Category	Basic Items	Evolving Items	Unlocked
facialHairType	Blank, BeardMedium, BeardLight, MoustacheFancy, MoustacheMagnum		
skinColor	Tanned, Yellow, Pale, Light, Brown, DarkBrown, Black		
eyeType	Default	Close, Cry, Dizzy, EyeRoll, Happy, Hearts, Side, Squint, Surprised, Wink, WinkWacky	Level 1
eyebrowType	Default	Angry, AngryNatural, DefaultNatural, FlatNatural, RaisedExcited, RaisedExcitedNatural, SadConcerned, SadConcernedNatural, UnibrowNatural, UpDown, UpDownNatural	Level 2
topType - hair	NoHair, LongHairBigHair, LongHairBob, LongHairCurly, LongHairDreads, LongHairFro, LongHairStraight, ShortHairDreads01, ShortHairShortCurly, ShortHairShortFlat, ShortHairSides, ShortHairTheCaesar	LongHairBun, LongHairCurvy, LongHairFrida, LongHairFroBand, LongHairNotTooLong, LongHairShavedSides, LongHairMiaWallace, LongHairStraight2, LongHairStraightStrand, ShortHairDreads02, ShortHairFrizzle, ShortHairShaggyMullet, ShortHairShortRound, ShortHairShortWaved, ShortHairTheCaesarSidePart	Level 3
mouthType	Default, Disbelief, Sad, Serious	Concerned, Eating, Grimace, ScreamOpen, Smile, Tongue, Twinkle, Vomit	Level 4
topType - top	Hijab, Turban	Eyepatch, Hat, WinterHat1, WinterHat2, WinterHat3, WinterHat4	Level 5
hairColor	Black, Blonde, Brown	Auburn, BlondeGolden, BrownDark, PastelPink, Platinum, Red, SilverGray	Level 6
topColor		Black, Blue01, Blue02, Blue03, Gray01, Gray02, Heather, PastelBlue, PastelGreen, PastelOrange, PastelRed, PastelYellow, Pink, Red, White	Level 7
facialHairColor	Black, Blonde, Brown	Auburn, BlondeGolden, BrownDark, PastelPink, Platinum, Red, SilverGray	Level 8
accessoriesType		Blank, Kurt, Prescription01, Prescription02, Round, Sunglasses, Wayfarers	Level 9
graphicType		Bat, Cumbia, Deer, Diamond, Hola, Pizza, Resist, Selena, Bear, SkullOutline, Skull	Level 10

response. Furthermore, Table 5 shows an overview of the descriptive statistics of our qualitative data analysis.

B.4 Task Differences

The task differences in perceived workload, intrinsic motivation, and subjective user engagement between the credibility analysis task and the information finding task can be found in Figure 14.

B.5 GIM level differences

The details of the exploratory statistic analyses for the Dunn tests between the different levels of group identification can be found in Table 6 (information finding task) and Table 7 (credibility analysis task).

Table 3: Descriptive statistics for the avatar editor and worker community space. The number of changes describes how often a worker changed features to edit their avatar. Total interactions describes how many times a worker clicked one of the interactive buttons in the worker community space, and the Time (in seconds) describes the amount of time the worker spent in the worker community space.

			Information Finding task			Credibility Analysis task		
	Measurement	**Condition**	Median	Mean	SD	Median	Mean	SD
Avatar Editor	Number of Changes	**Basic**	9	12.85	10.22	9	9.46	6.73
		Basic⊕Comm	9	11.69	8.81	9	13.40	14.67
		Evolving	17	20.40	17.48	16	29.75	45.53
		Evolving⊕Comm	14	19.69	18.70	13.5	26.84	29.38
Community Space	Total Interactions	**Basic⊕Comm**	0	0.87	1.25	0.5	0.85	1.23
		Evolving⊕Comm	0	0.74	1.24	1	0.91	1.26
	Time (s)	**Basic⊕Comm**	20.34	23.83	15.41	21.60	29.06	27.21
		Evolving⊕Comm	19.67	25.26	29.85	21.86	33.52	64.54

Table 4: Categories emerging from the open-coding of responses from the open-ended question on why the workers did or did not feel connected to other workers who completed the same tasks. The categories are described, and an example from the open-ended responses is presented as it stands in the original quotes. Furthermore, the quote in the title of this paper is an adjusted version of the original quote from our data marked in this table with *.

Category	Description	Example Response
Fake	Worker felt as if the avatars on the worker community page were not real workers.	*They were just icons on my screen and did not feel like real people.*
Feelings -	Seeing the feelings of other workers about the task made the worker feel less connected.	*People had different feelings.*
Feelings +	Seeing the feelings of other workers about the task made the worker feel more connected.	*Most of the other workers were relaxed and calm just like me.*
Interaction	The worker experienced a lack of interaction/ the worker mentions working solely/ the worker does not know other workers personally.	*I just saw them at the end. During the experiment there was no interaction.*
Avatar -	Seeing the avatars did not make the worker feel connected.	*It's hard to feel connected to someone behind an avatar with very little customisation.*
Avatar +	Seeing the avatars made the worker feel connected.	*The last page made me feel connected because we were all shown together.*
Shared Goal	Worker feels connected because they work on the same task.	**We're all in the same boat, doing the same thing for the same compensation.*
Other	All answers that did not fit the categories.	*I just didn't feel any connection.*

Table 5: The number (N) and percentage (%) of workers for each category of why they felt not connected (Connected <4) or connected (Connected >4) for both tasks. The total row describes the number and percentage of workers per task who felt connected or not.

	Information Finding task				Credibility Analysis task			
	Not connected		*Connected*		*Not connected*		*Connected*	
Category	N	%	N	%	N	%	N	%
Fake	8	13%	0	0%	2	3%	0	0%
Feelings -	4	6%	0	0%	4	6%	0	0%
Feelings +	0	0%	10	23%	1	2%	10	20%
Interaction	41	65%	3	7%	48	76%	7	14%
Avatar -	3	5%	1	2%	4	6%	0	0%
Avatar +	1	2%	4	9%	2	3%	4	8%
Shared Goal	1	2%	12	28%	2	3%	18	36%
Other	16	25%	11	26%	8	13%	14	28%
Total	63	47%	43	32%	63	46%	50	37%

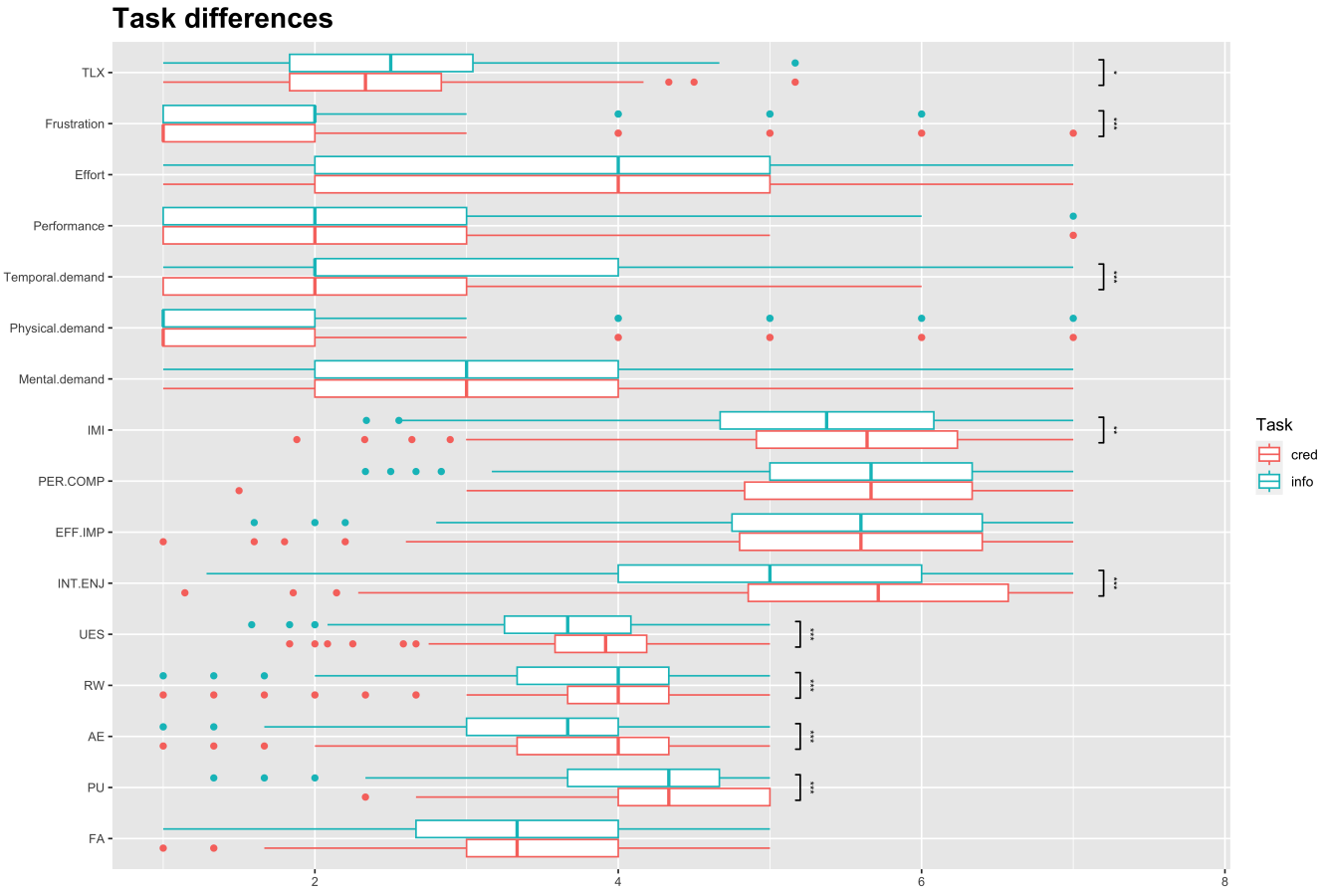


Figure 14: Significant differences between the worker experience of the credibility analysis task (*cred*) and the information finding task (*info*). * means $p < .05$, ** means $p < .01$, and * means $p < .001$. The TLX and IMI scores are measured on a 7-point Likert scale, and the UES measurements are measured on a 5-point Likert scale. Note that for the TLX measurements, a low score for the subdimension performance indicates a high perceived performance.**

Table 6: Results for the Dunn test for significant differences between different levels of group identification (GIM: low, mid, and high) within worker experience measurements for the information finding task. * means $p < 0.05$, ** means $p < 0.01$, and * means $p < 0.001$.**

	Information Finding task			
Measurement	Comparison	Z	p_unadj	p_adj
FA	High - Low	5.382	<.001	<.001***
	High - Mid	1.138	.255	.765
	Low - Mid	-4	<.001	<.001***
PU	High - Low	1.505	.132	.397
	High - Mid	3.479	.001	.002**
	Low - Mid	1.882	.06	.18
AE	High - Low	7.599	<.001	<.001***
	High - Mid	3.246	.001	.004**
	Low - Mid	-4.093	<.001	<.001***
RW	High - Low	8.341	<.001	<.001***
	High - Mid	4.037	<.001	<.001***
	Low - Mid	-4.042	<.001	<.001***
UES	High - Low	7.83	<.001	<.001***
	High - Mid	3.646	<.001	.001**
	Low - Mid	-3.931	<.001	<.001***
INT-ENJ	High - Low	8.572	<.001	<.001***
	High - Mid	3.72	<.001	.001**
	Low - Mid	-4.561	<.001	<.001***
EFF-IMP	High - Low	8.102	<.001	<.001***
	High - Mid	3.865	<.001	<.001***
	Low - Mid	-3.98	<.001	<.001***
PER-COMP	High - Low	5.306	<.001	<.001***
	High - Mid	4.28	<.001	<.001***
	Low - Mid	-0.947	.344	1
IMI	High - Low	9.023	<.001	<.001***
	High - Mid	4.729	<.001	<.001***
	Low - Mid	-4.029	<.001	<.001***
Mental demand	High - Low	3.867	<.001	<.001***
	High - Mid	1.097	.273	.818
	Low - Mid	-2.609	.009	.027*
Physical demand	High - Low	3.644	<.001	.001**
	High - Mid	1.246	.213	.638
	Low - Mid	-2.257	.024	.072
Effort	High - Low	5.243	<.001	<.001***
	High - Mid	2.578	.01	.03*
	Low - Mid	-2.503	.012	.037*
Frustration	High - Low	-1.937	.053	.158
	High - Mid	-2.771	.006	.017*
	Low - Mid	-0.801	.423	1
TLX	High - Low	3.097	.002	0.006**
	High - Mid	0.019	.985	1
	Low - Mid	-2.906	.004	.01*

Table 7: Results for the Dunn test for significant differences between different levels of group identification (GIM: low, mid, and high) within worker experience measurements for the credibility analysis task. * means $p < 0.05$, ** means $p < 0.01$, and * means $p < 0.001$.**

	Credibility Analysis task			
Measurement	Comparison	Z	p_unadj	p_adj
FA	High - Low	2.76	.006	.017*
	High - Mid	0.165	.869	1
	Low - Mid	-2.356	.018	.055
AE	High - Low	6.719	<.001	<.001***
	High - Mid	4.239	<.001	<.001***
	Low - Mid	-2.049	.04	.121
RW	High - Low	5.669	<.001	<.001***
	High - Mid	3.317	.001	.003**
	Low - Mid	-1.979	.048	.143
UES	High - Low	4.813	<.001	<.001***
	High - Mid	2.137	.033	.098
	Low - Mid	-2.331	.02	.059
INT-ENJ	High - Low	6.819	<.001	<.001***
	High - Mid	2.943	.003	.01*
	Low - Mid	-3.386	.001	.002**
EFF-IMP	High - Low	7.087	<.001	<.001***
	High - Mid	2.753	.006	.018*
	Low - Mid	-3.812	<.001	<.001***
PER-COMP	High - Low	6.502	<.001	<.001***
	High - Mid	2.694	.007	.021*
	Low - Mid	-3.335	.001	.003**
IMI	High - Low	7.914	<.001	<.001***
	High - Mid	3.29	.001	.003**
	Low - Mid	-4.05	<.001	<.001***
Mental demand	High - Low	2.465	.014	.041*
	High - Mid	0.574	.566	1
	Low - Mid	-1.694	.09	.271
Physical demand	High - Low	3.131	.002	.005**
	High - Mid	2.276	.023	.068
	Low - Mid	-0.666	.505	1
Effort	High - Low	3.171	.002	.005**
	High - Mid	1.83	.067	.202
	Low - Mid	-1.132	.258	.773
TLX	High - Low	2.757	.006	.018*
	High - Mid	2.025	.043	.129
	Low - Mid	-0.567	.571	1