Making the Switch: Towards Intelligent Integration of Gestures As an Input Modality for Microtask Crowdsourcing

Garrett Allen
Delft University of Technology
Delft, Netherlands
g.m.allen@tudelft.nl

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

Abstract

Human input is pivotal in building AI systems. Aiding the gathering of high-quality and representative human input on demand, microtask crowdsourcing platforms have thrived. Despite the benefits available, the lack of health provisions, safeguards, and existing practices threaten the sustainability of crowd work. Prior work investigated the usefulness of a dual-purpose input modality of ergonomically-informed gestures across different microtasks, finding that gestures as inputs offer a realistic trade-off between worker accuracy and potential short to long-term health benefits. However, little is understood about the effect of switching input modalities from one task to another on worker experiences and task-related outcomes. Addressing this research and empirical gap, we conducted a between-subjects study ($N = 717$) with varying sequences of input modalities across 16 experimental conditions to systematically understand the effect of switching input modalities. We found that the order of the input modality can influence the time it takes to complete tasks but does not affect accuracy. Further, the cognitive load perceived by workers was not significantly different between conditions. Our findings hint that ergonomically informed gestures can be effectively intertwined with conventional input modalities without a detrimental impact on worker experiences and quality-related outcomes. Our work has important implications for the design of human-centered crowdsourcing platforms that cater to worker health and wellbeing.

ACM Reference Format:
Garrett Allen and Ujwal Gadiraju. 2025. Making the Switch: Towards Intelligent Integration of Gestures As an Input Modality for Microtask Crowdsourcing. In *CHIWORK '25: Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work (CHIWORK '25), June 23–25, 2025, Amsterdam, Netherlands.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3729176.3729184

1 Introduction

Crowdsourcing marketplaces provide a centralized place for requesters to post microtasks to gather cost-effective, high-quality data by harnessing human intelligence at scale [20, 24, 29]. This is still a growing and evolving paradigm with a multitude of unsolved challenges [4, 23, 55]. A vast majority of early research in this field has focused on understanding the use of crowdsourcing in realizing specific use cases [23] and tackling quality control challenges [11, 26]. Recent efforts have turned their attention to topics such as understanding and improving worker experiences and supporting workers [12], worker engagement [34, 42], and ensuring fair compensation of workers [6, 52], among others. Such worker-centric explorations are valuable in a world where microtask crowdsourcing centers around online marketplaces that are subject to evolving interests, and where the abuse of data workers continues to be prevalent [20, 32, 48].

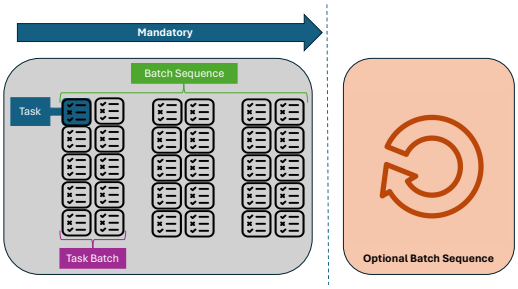


Figure 1: An illustration of the task batches in our controlled study. Workers completed a sequence of task batches, with each batch containing 10 tasks.

Conventional crowd work in various microtask marketplaces bears a resemblance to desk work in that the work is often conducted using the familiar mouse and keyboard input devices. Other input modalities that have been utilized for crowdsourcing studies include voice, eye tracking, and hand gestures [8, 30, 50, 54]. Recent work by Allen et al. [3] has proposed using ergonomically-informed gestures to help improve health-related aspects in crowd work. Workers on microtask platforms complete tasks in sequences — either within or across different task batches. Newell and Ruths [36] investigated *intertask effects*, and showed that earlier tasks in batches can have a large effect on later tasks. Cai et al. [7] explored the effects of task complexity on performance and worker experience. Within the frame of writing tasks, i.e., producing a written passage of text, the authors explored three distinct perspectives: *continuity* – task chains of the same complexity, *transition* – moving between tasks of different complexity within a chain, and *easing in* – moving from lower complexity to higher complexity tasks in a chain. Taking inspiration from this, Aipe and Gadiraju [2] explored the effects of task *similarity*, i.e., the degree of resemblance between a pair of tasks. These works found that task complexity and similarity shape task-related outcomes. In this work, we aim to extend the understanding of task effects on workers by delving into the currently unexplored impact of varying *input modality* across

CHIWORK '25, Amsterdam, Netherlands

ACM ISBN 979-8-4007-1384-2/25/06
https://doi.org/10.1145/3729176.3729184

a sequence of task batches. Figure 1 illustrates the differences between a task, a batch, and a sequence. A task is an atomic entity consisting of a single prompt. A batch is a collection of ten tasks, and a batch sequence is a series of three batches. We also aim to further the understanding of gesture-based input for crowd work by addressing the following research question:

> ***RQ:*** *How does integrating gesture-based input with conventional input modalities influence task-related outcomes?*

To address this research and empirical gap, we carried out a between-subjects study across 16 different experimental conditions with varying sequences of input modalities across two types of tasks. Drawing from prior work and accounting for popular microtask crowdsourcing tasks [13, 19], we considered a sentiment analysis (SA) task where workers are asked to assign star ratings to movie reviews based on the content of the review, and a classification task (CC) related to the shapes of bird beaks [3]. Our custom website integrates, *i.e.*, includes one or both, `Standard` and `Gesture` inputs into the task workflow workers perform, per the experimental condition. We further capture and report objective measures of effectiveness (task accuracy) and efficiency (task completion time).

Main Contributions. Our contributions include novel insights on the impact of integrating multiple input modalities within a batch sequence on task-related outcomes. Findings from our rigorous between-subjects study (*N=717*) with varying sequences of input modalities across 16 experimental conditions suggest that the order of the input modality sequence can influence the task completion time but does not affect the average accuracy. We contribute a clearer understanding of worker perceptions around gestures as inputs for microtasks, such as a lack of significant difference in cognitive load perceived by the workers between different conditions. Our findings suggest that ergonomically informed gestures can be effectively intertwined with conventional input modalities without negatively impacting worker experiences. We found trends that indicate potential trade-offs with quality-related outcomes despite a lack of significant differences in worker performance across the experimental conditions on average. These insights have important implications for the future design of crowdsourcing marketplaces and data acquisition.

2 Related Literature and Hypotheses

We position our findings and contributions in (a) intelligent task chaining in microtask crowdsourcing, (b) input modalities, and (c) worker experiences in crowdsourcing. We present our hypotheses and the rationale behind each.

2.1 Intelligent Task Chaining in Crowdsourcing

In microtask crowdsourcing, workers do not typically complete only a single task at a time. Instead, they gather tasks and perform them in batches [53]. Similarly, multiple tasks can be completed in parallel or sequence within workflows. Newell and Ruths [36] investigated what the authors call *intertask effects*, i.e., the effect that one task has on how workers perform subsequent tasks. Earlier tasks within a chain were found to have a strong influence over the following tasks. At the same time, Cai et al. [7] explored microtask chains from the perspective of order, i.e., the authors focused on the effects of *transitions*, *ease-in*, and *continuity*. Transitions are the process of moving between tasks of varying complexity within a chain, and ease-in is the idea of moving from simpler microtasks to more complex ones. Continuity refers to performing a chain of tasks that are of similar complexity. By digging deeper into the facets of task similarity, Aipe and Gadiraju [2] found that accuracy improves if tasks in succession are similar. In this paper, we build on this existing body of work by exploring the impact of modality-based task chains on worker experiences and task outcomes.

2.2 Common Input Modalities in Crowdsourcing

Input modalities of the mouse and keyboard are ubiquitous, while others such as eye-tracking-based triggers are more niche in crowdsourcing. *Gestures*, defined by Carfì and Mastrogiovanni [8] as "body actions that humans intentionally perform to affect the behavior of an intelligent system" are an alternative input modality. Prior work has separated gestures into "vision-based" methods that rely on computer vision strategies to interpret gestures as input [28, 51] and "sensor-based" methods that rely on tangible, often wearable interfaces [1]. Quek [43] separates gestures into "communicative", i.e., used to communicate, and "manipulative" or used to interact with objects or systems for effect. In their seminal work, Pekin [40] provided a library of gestures suitable for large screen-size touch interfaces. There are many variations of gestures for human-machine interaction, but none are explicitly defined for crowdsourcing microtasks. Allen et al. [3] investigated the viability of gestures as an alternative input modality within crowdsourcing, finding a trade-off between performance and perceived reward of the input modality. In our study, we aim to extend the understanding of gestures for microtask crowdsourcing by exploring the effect of gestures within a batch of tasks, as opposed to considering them in isolation which prior work has been limited to.

2.3 Worker Experiences in Crowdsourcing

The complex dynamics in crowdsourcing marketplaces between platforms, task requesters, and workers represents an ecosystem that is far from ideal [16, 17, 21, 31, 44, 48]. Over the years, several researchers and practitioners have addressed this myriad of issues with the aim to increase wages [52], improve practices [38, 47], help workers' growth [10], enhance worker experiences [42], engagement [27], and retention [12, 21]. Our work contributes to this wealth of literature by exploring the impact of integrating gestures as an input modality on worker experiences and engagement. Inspired by prior work, we explore the impact of varying input modality on the cognitive load perceived by workers, their subjective engagement, and their objective retention in task batches.

2.4 Hypotheses

Prior exploration of gesture inputs within microtask crowdsourcing found workers were more accurate in tasks completed with the `Standard` input compared to those with the `Gesture` input [3]. Workers have also been shown to perform more accurately in task batches composed of similar tasks, despite perceived boredom [2]. Thus, we extrapolate that workers may perform better in task

batches with a single input modality compared to multiple input modalities.

We also anticipate that switching modalities between batches of tasks of the same purpose may offset the potential for boredom. As a result, workers will perceive and demonstrate a higher engagement with the tasks. This informs the following hypotheses (**H1**, **H2**):

> **H1**: Workers will complete tasks more accurately when the input modality remains constant within a task batch.
> **H2**: Worker engagement will increase when task batches include different input modalities.

Allen et al. [3] found that workers experience a higher cognitive load while using gesture inputs. Thus, we expect that the context of switching from a conventional mouse and keyboard input (`Standard`) to that of gestures (`Gesture`) would increase the cognitive effort of workers. In the same vein, switching from `Gesture` to `Standard` would decrease their perceived cognitive effort. Finally, drawing from prior work we expect that switching input modalities within a task batch will lead to a longer task completion time. This informs the following (**H3**, **H4**, **H5**):

> **H3**: The perceived cognitive effort of workers will increase when switching from tasks using a `Standard` input to tasks using a `Gesture` input modality.
> **H4**: The perceived cognitive effort of workers will decrease when switching from tasks using a `Gesture` input to a `Standard` input modality.
> **H5**: Workers will take longer to complete task batches with varying input modalities than when using a single modality.

3 Method

To address our research question, test the aforementioned hypotheses, and understand the effects of switching input modalities within a batch of similar tasks, we designed and carried out a controlled 8×2 factorial between-subjects study. The independent variables are (i) the sequence of transitions (or switches) across *input modalities* and (ii) the *task type*. We consider two different input modalities — a mouse and keyboard (`Standard`) and webcam-based gesture capture (`Gesture`). We configure the order of these modalities into eight variations to emulate all possible sequences in a batch of tasks: *SSG*, *SGS*, *GGS*, *GSS*, *SGG*, *GSG*, *GGG*, and *SSS* where "`S`" represents the `Standard` input and "`G`" represents `Gesture` input. To capture whether the task type itself influences the effects of switching input modality within a batch, we also utilize two task types from the taxonomy in [19]: *sentiment analysis* (SA) and *categorization and classification* (CC). We selected these task types based on similar studies [3] and their prevalence in online crowdsourcing platforms.

Sentiment Analysis. The sentiment analysis task involves reading a movie or television program review and assigning a star rating on a scale from 1 to 5. We sampled sixty reviews from the Amazon Review Dataset [37], with a balanced distribution of star ratings, i.e., twelve of each possible rating. For consistency, we limited the selection of reviews to ones with a total length of 150 words. Depending on the experimental condition, participants provide the rating in one of two ways, with a `Standard` or `Gesture` interface. The `Standard` interface, as seen in Figure 2(a), presents the text of a review in the center, with selectable star icons below the text. A button for submitting an answer is initially greyed out (inactive) until a star rating is selected. The `Gesture` interface (Figure 2(c)) is identical to that of the `Standard`, with the addition of a camera view below the stars through which the gesture interaction takes place. This camera view is the driver for the gesture capture, with the vertical red bars dividing the view into sections that map to a star rating; one star at the far left up to five stars at the far right. Moving a closed fist between partitions allows workers to change the rating without submitting it as an answer. By opening their palm, the counter seen below the camera view activates. This timer starts at two and counts down to zero. Once it reaches zero, a selection is made and submitted as the answer, and the next review to rate in the task batch is displayed.

Categorization and Classification. In this task, workers are asked to analyze the image of a bird and identify the shape of the bird's beak. For the images, we use the dataset introduced by Balayn et al. [5], which includes images of eight different beak types. In total, the dataset contains 79 images. As with the *SA* tasks, sixty images were selected and workers performed the classification with either the `Standard` or `Gesture` input. When using the `Standard` input modality, the image being classified is displayed in the top center of the page, as seen in Figure 2(b). A representative of each beak shape is displayed below the main image, along with their name, as selectable answers.

For the `Gesture` interface, the camera view has only a single, horizontal red line (Figure 2(d)) instead of the five vertical lines in the *SA* task. Workers are required to raise the number of fingers (i.e., digits) corresponding to the answer option they wish to select. To optimally leverage the gesture capture model, thumbs are ignored as a digit. While keeping their hands below the horizontal line, workers can change their selection by varying the number of digits at will. To submit, workers must raise their hands with presented digits above this threshold, initializing the counter. Once the counter reaches zero, the selected option is submitted as an answer, and the next task is presented to the workers.

3.1 Procedure

We deployed our study on Prolific, where willing and eligible workers were directed to the website containing the embedded *Qualtrics* pre-task and post-task surveys[1] as well as the custom task interface. Upon entering the external site, workers completed the following five steps:

(1) **Informed Consent**: Before performing the task or interacting with any surveys, workers are presented with an informed consent form outlining the study's purpose and what they could expect in the study.
(2) **Pre-task Questionnaire**: A questionnaire focused on demographic questions such as age, experience on Prolific, mood, and time spent on crowdsourcing work.
(3) **Training Phase**: Workers were required to complete a sample of questions from each task type and input modality that would

[1] https://qualtrics.com

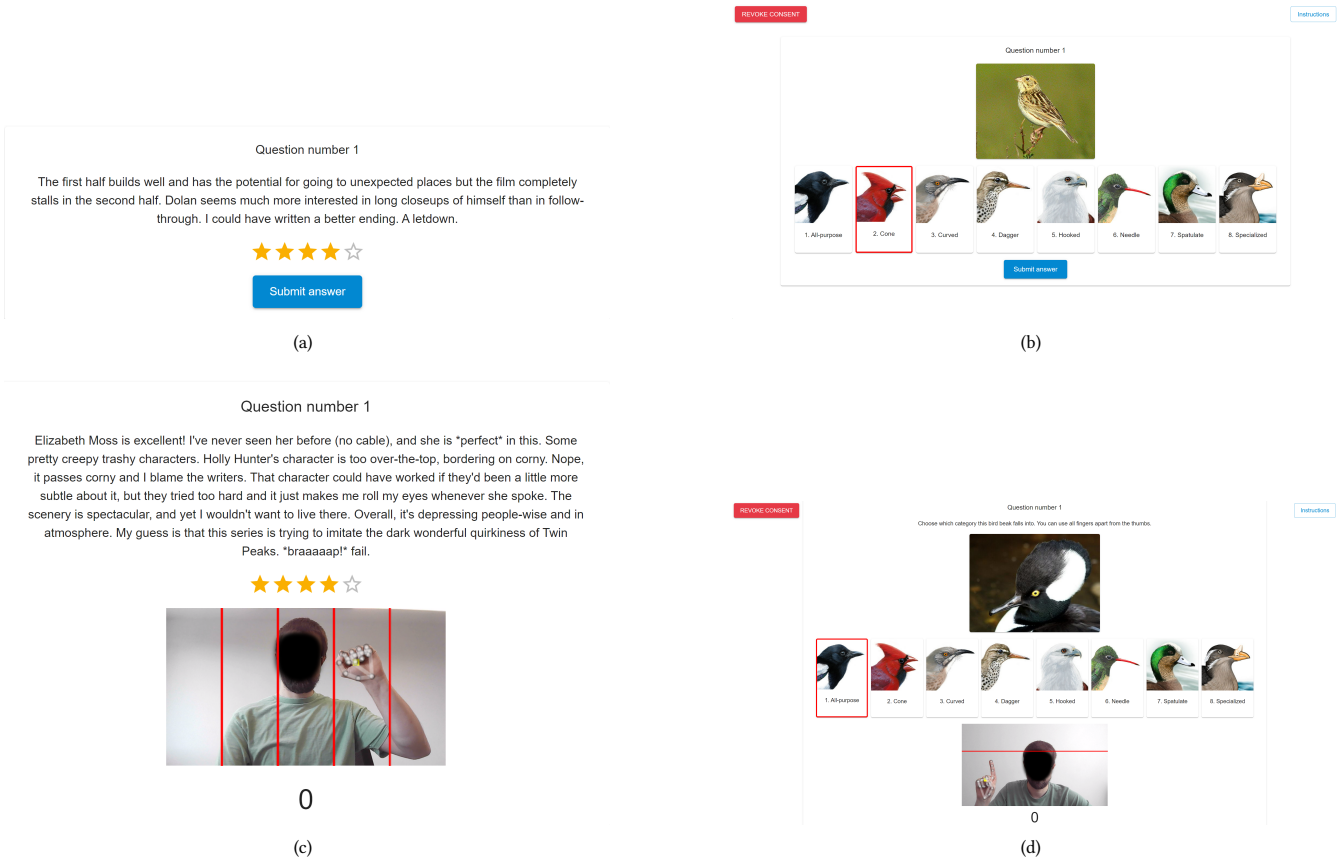


Figure 2: Example screenshots of task interfaces: (a) Sentiment Analysis with `Standard` input. (b) Classification with `Standard` input. Second row, left to right: (c) Sentiment with `Gesture` input. (d) Classification with `Gesture` input.

be utilized in their respective experimental conditions. This training phase aims to allow workers to familiarize themselves with the task and the corresponding input modality before completing the actual tasks; and also minimizes potential familiarity bias [14].

(4) **Batch Sequence**: Each sequence contains a batch of sentiment analysis or content classification tasks depending on the experimental conditions. Each batch is ten tasks with either a `Standard` or `Gesture` input. The order of inputs in the batch sequence is defined by the experimental conditions. After completing the batch sequence, workers can choose to continue and complete up to three more optional task batches.

(5) **Post-task Questionnaire**: A follow-up questionnaire presented after the main batch sequence to capture measures necessary to test our hypotheses.

When workers complete the post-task questionnaire, they are provided a completion code and redirected back to Prolific.

3.2 Technical Implementation

Inspired by prior work by Allen et al. [3], all experimental conditions were hosted on a custom-developed website built using the ReactJS library. This website has three major components: a back-end server and API, a front-end web application, and a MongoDB NoSQL database. The back-end server manages each worker's progress by processing requests sent by the front end via an HTTP REST API. This server also handles communication with the database, where all worker data is stored. The front-end client renders the pages and handles gesture, mouse, and keyboard inputs. Two consecutive client-side stages enable the input of gestures via the webcam of a participant: pose detection and pose classification.

Gesture Capture with Pose Detection and Classification. Considering participant privacy as a priority, we approached the design and development of the gesture capture to be a client-side system, i.e., powered by the device of the crowd worker. We used the same libraries as were included in the work by Allen et al. [3]. The MediaPipe holistic[2] and Kalidokit[3] libraries handle processing the webcam video feed via pre-trained models. The models perform landmark estimation for the body, face, and hands visible in the webcam feed, which is then used to classify the poses.

3.3 Participants and Measures

We computed the required sample size in a power analysis for a Between-Subjects ANOVA using the G*Power software. Assuming

[2] https://google.github.io/mediapipe/solutions/holistic.html
[3] https://github.com/yeemachine/kalidokit

$f = 0.25$, a significance threshold of $\alpha = \frac{0.05}{5} = 0.01$ (due to testing five hypotheses), a power of (1- β) = 0.95 and given that we will be testing 16 groups (i.e., eight batch sequences with different input modality conditions, two task types), we determined via a power analysis for one-way ANOVA (see Section 4) that the required sample size for our study is 735 participants.

Participants were recruited from the online crowdsourcing platform *Prolific* and were required to be fluent English speakers above 18 years of age. To ensure quality responses, we limited participation to those with a minimum 90% approval rate. As our study required the capturing of webcam-based gestures, participants needed a webcam. Each participant was allowed to enter our study only once by leveraging in-built functions on the platform.

User Engagement. To understand the impact of switching input modalities on workers' engagement, we used the User Engagement Scale-Short Form (UES-SF). This validated questionnaire includes four sub-scales to capture engagement – focused attention, perceived usability, aesthetic appeal, and reward [39]. Each sub-scale contains three questions that workers respond to using a 5-point Likert scale. An engagement score is determined as an average score across all sub-scales. To capture an objective notion of engagement, we used *worker retention* as a measure of user engagement [12, 45]. Workers can continue with optional task batches within each experimental condition, completing a similar batch sequence as in the mandatory segment. For example, a participant in the condition *SSG + SA* could complete an optional segment that followed the same sequence of input modality and task type.

Cognitive Load. To gain insight into whether switching modalities during a batch sequence impacts the perceived cognitive load of workers, we use the NASA-TLX [22]. An overall cognitive load score is determined as an average of the worker's responses across the six questions captured on a 10-point scale (we used an unweighted score).

Completion Time. We record timestamps (captured in milliseconds) when workers begin and end each task.

Accuracy. The accuracy of a worker (i.e., their task performance) is determined by the fraction of tasks for which they provided an answer matching a known ground truth.

3.4 Data Privacy

To protect the workers' privacy, we did not collect any personally identifiable information. All data is only linked to the Prolific IDs of workers during data collection, including survey responses, question answers, and pose data. Upon completion of data collection, all data is assigned a unique identifier, and the Prolific IDs are discarded to prevent back identification. Webcam images used for gesture recognition are shown and processed on the participants' devices and never sent to the back end or stored anywhere. Instead, we collect pose landmarks on some actions, e.g., whether a pose was started or ended, but these do not constitute personally identifiable information. Workers can revoke their consent at any time during data collection and any data generated from such workers is permanently removed. The research and informed consent materials received approval from an institutional Human Research Ethics Committee.

4 Results

All statistical analyses were performed using **JASP**.[4] Data and results are available in an OSF repository.[5] Results presented in this section emerge from corresponding ANOVA tests. Hypothesis testing is conducted with an adjusted significance threshold of $\alpha = \frac{0.05}{5} = 0.01$. The Bonferroni correction was applied with an original target of Type I error probability being $\alpha = 0.05$. Type II errors are accounted for using the Bonferroni-Holm correction after p-values are attained from the initial ANOVA tests. We conducted post hoc analysis via Tukey's HSD test for significant results after corrections. Workers were excluded from the study and our analysis if they failed two or more attention checks, revoked consent, provided the same answer across all Likert scale questions, and/or spent less than a minimum of fifteen minutes to complete the study. In sum, there were 84 exclusions.

4.1 Worker Demographics

After exclusions, we recruited 717 workers from Prolific. Workers recruited per experimental condition can be seen in Table 1. Complete details of worker demographics are illustrated in the OSF repository. The majority of workers were female ($n = 375$), followed by male ($n = 330$), with nine reporting as non-binary, one as other, and two preferring not to say. Most of the workers had 1–3 years of crowd work experience while performing tasks <10 hours per week. Workers ranged in age from 18 to 75 years and worked at a variety of times during the day.

4.2 Task Performance

We measure worker performance via accuracy, separating the required and optional tasks. Using an ANOVA, we found evidence indicating differences between the batch sequences of task types, during both the required and optional portions of the study (see Table 2), but no impact of the input modality on accuracy. A post hoc Tukey's HSD test indicated that workers on average were significantly more accurate at performing the `CC` task than the `SA` task ($p < 0.001$). Therefore, we reject **H1**.

While our initial analysis indicated significant differences between batch sequences, the analysis was performed on the average accuracy across the batch sequences. To get a more complete indication of the worker performance, we further investigated performance by plotting the accuracy *trend* of the different task types after each task batch. When looking at these trends, an interesting pattern becomes apparent: within a batch sequence, the accuracy decreases whenever the input modality changes from `Standard` to `Gesture`, with some decreases being larger than others. This pattern is visible for both task types and is most apparent for the `SGS` input sequence (see Figures 3(b) & 3(a)). The opposite is also true, worker accuracy increases when the input modality switches from `Gesture` to `Standard`. Both trends indicate a clear impact on performance when modalities change within a batch sequence.

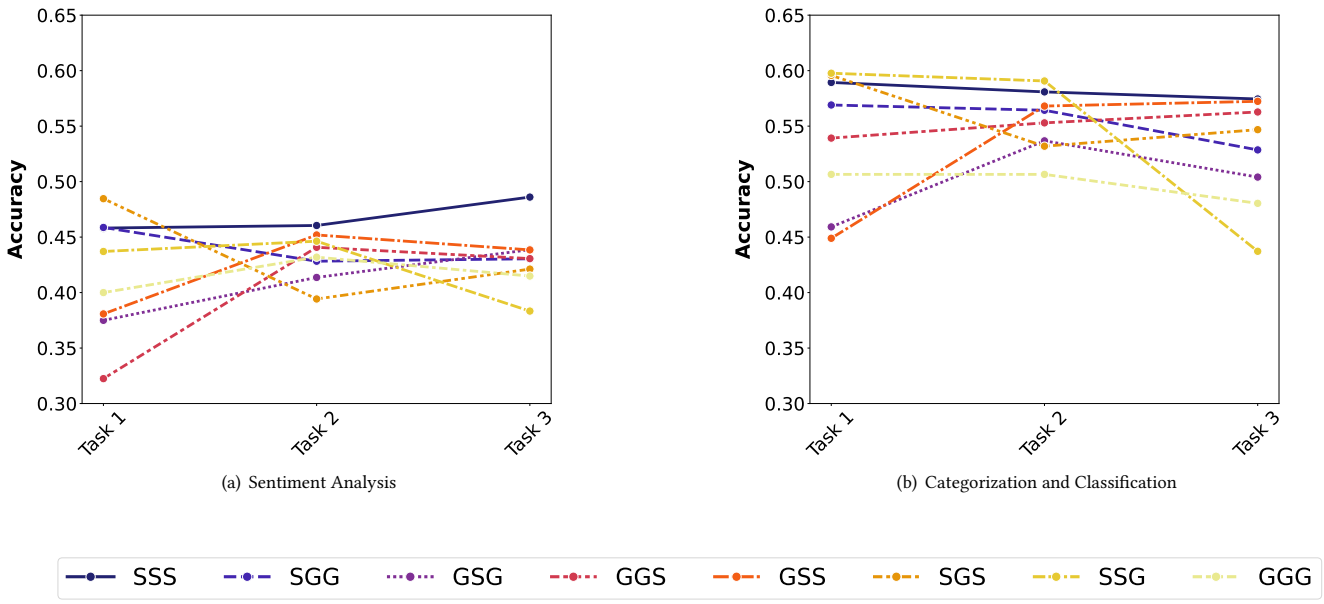
As part of the post-task questionnaire, we asked workers to report the percentage of tasks they believed to have completed correctly. Workers reported a significantly higher accuracy for the

[4] https://jasp-stats.org/

[5] https://osf.io/8rkeh/?view_only=eabf58aa53974195a4aa84c5834f454d

Task Type	Input Sequence	# Workers	Accuracy	Completion Time ± SD	Retention Rate	Avg. Depth
SA	SSG	53	0.43	164.1 ± 65.2	0.58	11.26
	SGS	51	0.43	192.4 ± 65.5	0.61	12.96
	GSS	44	0.41	192.9 ± 74.4	**0.66**	14.23
	GGS	46	0.42	**210.6 ± 99.9**	0.63	12.15
	GSG	43	0.44	200.5 ± 72.9	0.51	12
	SGG	42	0.45	193.3 ± 79.1	0.57	11.57
	SSS	*42*	***0.49***	*190.9 ± 78.4*	*0.62*	*12.21*
	GGG	*44*	*0.43*	*189.3 ± 74.6*	***0.66***	***15.39***
CC	SSG	41	0.55	181.7 ± 82.6	0.73	18.44
	SGS	46	0.58	201.6 ± 89.2	0.63	13.48
	GSS	44	0.55	204.7 ± 68.2	0.70	15.25
	GGS	49	0.55	**205.4 ± 118.6**	**0.80**	**19.8**
	GSG	43	0.53	191.7 ± 83.8	0.70	15.63
	SGG	39	0.55	204.5 ± 103.8	0.67	17.69
	SSS	*47*	***0.60***	*188.6 ± 117.8*	*0.72*	*15.74*
	GGG	*44*	*0.50*	*190.5 ± 77.2*	*0.60*	*14.63*

Table 1: Number of workers, retention rate, average depth (i.e., number of extra questions answered), and the mean completion time per experimental condition. We use bold font to indicate the highest value per task type and *italics* to indicate the control conditions that correspond to a single input modality — either `Standard` (i.e., SSS) or `Gesture` (i.e., GGG).



(a) Sentiment Analysis

(b) Categorization and Classification

Figure 3: Plot (a) shows the accuracy trend across the three tasks in the Sentiment Analysis batch. Plot (b) shows the same for the Categorization and Classification batch. The legend is shared across plots. *SSS* and *GGG* input sequences are controls.

SA (*M=0.82, SD=0.18*) compared to the CC (*M=0.73, SD=0.16*); see Table 3. This is counter to workers' actual performance – they performed more accurately in the CC tasks than the SA tasks. This can be attributed to the general familiarity that workers may have with reviews in contrast to identifying the shapes of bird beaks, potentially leading to crowd workers having less confidence in their assessments of the latter. An alternative explanation for this can be the Dunning-Kruger Effect, a meta-cognitive bias shown to be present and prominent in crowdsourcing marketplaces due to which workers exhibit an inflated self-assessment [14, 18]. We also asked workers about the task interface and whether it could result in errors. We found neither significant main effects for the batch or input sequences nor any significant interaction effect.

Variables	Required		Optional	
	F	p	F	p
Input Sequence	1.91	0.07	0.40	0.9
Batch Sequence	84.2	$<0.001^{\dagger}$	191.79	$<0.001^{\dagger}$
Interaction	0.63	0.73	1.15	0.33

Table 2: ANOVA test results for hypothesis H1 on worker accuracy. "Required" refers to the batch sequence workers must perform. "Optional" refers to the extra sequence workers can opt to complete. "Interaction" refers to the interaction effect between batch and input sequence. † indicates significance with a threshold of $p < 0.01$.

Variables	F	p
Input Sequence	2.844	$0.006^{\dagger}$
Batch Sequence	57.651	$<0.001^{\dagger}$
Interaction	2.978	$0.004^{\dagger}$

Table 3: ANOVA test results for worker perception of their performance. "Interaction" refers to the interaction effect between task and input sequence. † indicates significance with a threshold of $p < 0.01$.

Variables	Perceived Usability		Reward	
	F	p	F	p
Input Sequence	0.12	0.73	0.87	0.35
Batch Sequence	1.98	0.056	1.95	0.059
Interaction	2.93	$0.005^{\dagger}$	0.22	0.98

Table 4: ANOVA test results for H2 on worker engagement. "Interaction" refers to the interaction effect between task and input sequence. † indicates significance with a threshold of $p < 0.01$.

4.3 Worker Engagement

Analysis of worker responses to the UES-SF indicates there is no significant main effect for *perceived usability* for the input or batch sequences (see Table 4). However, there is a significant interaction effect between the batch and input sequences. A post hoc Tukey's HSD test uncovered that workers within the SSS+SA task perceived the interface as significantly more usable than those in the GGS+SA conditions. This difference may be related to the SSS+SA condition containing homogeneous inputs across all task batches, whereas GGS+SA contains an input modality transition.

Upon concluding the required batch sequence, workers were presented with the option to continue if they desired. Across all conditions, an average of 64.8% of workers opted to complete additional task batches. From Table 1, we can see that the CC task type resulted in higher retention rates across the conditions along with further retention depth. We measure the retention depth by the number of tasks beyond the required 30 that workers completed. The GGS+CC condition experienced the highest retention rate, with four out of five workers opting to continue. Those who continued completed an average of 19.8 additional tasks out of 30 (see the Avg. Depth column in Table 1).

Summary. Based on our results, we do not find evidence to support hypothesis **H2** that worker engagement increases when batch sequences include different input modalities (from the standpoint of both subjective and objective measures of worker engagement).

4.4 Cognitive Load and Task Completion Time

We hypothesized that the cognitive effort perceived by workers would increase when switching from tasks using a Standard input to tasks using Gesture input (**H3**) and would decrease when switching from Gesture to Standard (**H4**). Due to the directional nature of each hypothesis, we use ANOVA on the experimental conditions that contain only the transitions in the appropriate direction, i.e., we drop the SGS and GSG sequences. We explored the six NASA-TLX factors of mental demands, physical demands, temporal demands, own performance, effort, and frustration perceived by workers. We found no main effects for batch or input sequence, nor interaction effects, for the temporal demand, effort, and frustration factors of cognitive load. However, workers reported significantly better performance for the SA task. Further, we found a main effect concerning the input sequence for the physical demand factor; *$F(7, 701)=6.98$, $p<0.001$.* A post hoc Tukey's HSD test indicated that all input sequences corresponded to significantly higher physical demand in comparison to the SSS sequence (with $p < .001$ for all). This can be intuitively explained by the presence of gestures in every other input sequence, which requires relatively more physical effort than using mouse clicks and key presses.

Workers also reported the CC task type as significantly more mentally demanding than the SA task; *$F(1, 701)=7.51$, $p=0.006$.* We attribute this to the individual characteristics of the tasks. Reviews are common across the online landscape, connecting to things beyond movies, e.g., restaurants and shopping sites. The near-ubiquitous presence of reviews means that, intentionally or not, workers have passively garnered skills for their assessment. In contrast, such a passive development of skill is not as common for assessing the beak shape of birds. As a result, workers put forth more mental effort when analyzing the images of birds to determine beak shapes. The difference could also be due to the larger decision space for the CC task where workers must choose between eight different options.

When all of the factors are taken in conjunction as a single score representing aggregate cognitive load, we found that workers experienced no significant difference in cognitive load in the batch sequence or the input modality sequence.

With respect to task completion time, we found no main effects for batch or input sequence nor interaction effects in our analysis (see Table 1), and therefore reject **H5**. Interestingly, we found that the input sequence GGS took the most time for workers to complete, while the SSG sequence took the least time for both the SA and CC tasks.

Summary. We did not find support for our hypotheses **H3** and **H4** that worker cognitive load changes when switching between task batches with different input modalities. We also did not find support for hypothesis **H5** that workers take longer to complete batch sequences with varying modalities than using a single modality.

5 Discussion, Conclusions, and Future Work

Through the lens of sustainable crowd work and to foster inclusive crowdsourcing marketplaces, it is valuable to explore how barriers to participation can be lowered and how novel input modalities can be supported and adopted. In this spirit, Singh et al. [46] built a system called '*SignUpCrowd*' to support sign language as an input modality for crowdsourcing. Others explored how gestures as an input modality compare to conventional alternatives [3]. However, it is important to understand how alternative input modalities such as gestures can be integrated into existing marketplaces where tasks are not completed in isolation but in myriad sequences. This contextual detail has not been considered by prior work, limiting the ecologically valid understanding of the suitability of gestures as an input modality for crowdsourcing. What are the potential costs and trade-offs to consider if crowdsourcing platforms or task requesters are to integrate such alternatives into the general fabric of existing work?

In our study consisting of two different types of tasks, we found that workers completing the sentiment analysis task reported the homogeneous input sequence with the `Standard` input modality as significantly more usable than the `GSG` input sequence. Across all input sequences, this pair is the only case of a significant difference. Yet this finding provides evidence for potential effects on user engagement when using gestures as inputs. Future investigation of whether our finding is an outlier or an indicator of larger effects is required. We also found that the average task accuracy of workers, their perceived cognitive load, and task completion time remained unaffected when different input modalities were used within a batch sequence compared to those with a homogeneous input modality (**H1**,**H3**—**H5**). These novel insights suggest the potential for gestures to be successfully integrated as an input modality for microtask crowdsourcing alongside the conventional mouse and keyboard input modality without negatively affecting outcomes in batch sequences.

We found that the accuracy of workers in the sentiment analysis task (`SA`) was lower than in the categorization task (`CC`). In the `SA` task, we found a notable increase in accuracy corresponding to the `GGS` input sequence, which is not observed in the `CC` task type. The accuracy of workers in the `GGS` input sequence increased across the batches while decreasing in the `SSG` input sequence across the batches, suggesting potential changes in accuracy in favor of the `Standard` input modality. These changes may adversely affect data quality in scenarios where workers perform tasks of the same type with different inputs. Microtask crowd workers select and complete tasks from many requesters, often one after the other. From one requester to another, our findings reveal worker performance may fluctuate if the input modalities vary. Over time these effects may flatten as skills and familiarity with such novel input modalities increase among workers. Further work is needed to explore this.

Based on the observations from our investigation, it appears that `Standard` inputs outperform `Gesture`. However, given the equally positive worker experiences with the gestures in our study, we highlight the potential in integrating `Gesture` as an input modality in microtask marketplaces — particularly from the standpoint of exploring how `Gesture` can expand the scope of participation in crowd work and how ergonomically informed `Gesture` can have a positive impact on worker health and wellbeing. Note that in this work we are not advocating for gesture inputs to replace current standard inputs. Instead, we sought to understand how the two input modalities can co-exist within crowdsourcing marketplaces and how these input modalities would shape task outcomes.

Collectively, our insights provide evidence supporting the creation and integration of gestures as an alternative input modality in crowdsourcing tasks, while suggesting potential for trade-offs in performance in specific sequences of transitions across input modalities. These findings have broad implications for crowdsourcing platform design and practices for task requesters and crowd workers. Crowdsourcing platforms can build scaffolds, tools, plugins, or fundamental infrastructure, and create meaningful pricing models to support the adoption and inclusion of gestures as an alternative input modality. Task requesters can change existing practices to systematically enhance the range of modalities via which input can be gathered from crowd workers. The intelligent integration of different input modalities in task batches (e.g., `Standard` alongside `Gesture`) can serve as a valuable canvas towards addressing complex challenges surrounding worker engagement to worker long-term health and wellbeing (for instance, by designing ergonomically-informed gestures).

5.1 Caveats, Limitations, and Other Considerations.

In this study, we did not consider changing task types alongside changing input modalities within task batches (i.e., all task batches comprised homogeneous task types). Future work can explore the impact of varying both input modality and task types within task batches. Although the pose detection and classification modules we used were highly accurate, potential improvements in such models can further enhance worker experiences. Gestures as an input modality can potentially pave ways to safeguard worker health and wellbeing, but they also bear implications on who can participate — due to the need for a webcam (although most devices used for microtask crowdsourcing these days are equipped with webcams). It is important to consider the benefits of integrating gestures as an input modality for crowd work but also the potential harm in fragmenting access.

The Mapping Problem. Understanding gestures as inputs holistically is a challenging prospect. We need to explore how users perceive an input modality and understand what tasks are well-suited to be completed with gesture input. Each side of this exploration—(a) understanding user experiences with different gestures, and (b) feasibility of using gestures for specific tasks—benefits from knowledge of the other. We refer to this challenge as '*the Mapping Problem.*' Future work should explore the creation of an inventory of gestures that can be mapped onto different tasks either atomically or through composite combinations of gestures. The performance variance we observed among workers when switching from standard to gesture input, hint at adaptation costs that could potentially be mitigated through intelligent onboarding, dynamic task guidance, or predictive interface adjustments. These are promising research directions for future work.

Further, various human factors can affect a person's ability to perform a gesture, e.g., range of motion. Understanding how inclusive and accessible gestures are is an important question for future research. In our study, we did not gather explicit data pertaining to the accuracy of the gesture recognition or worker perceptions of the same. Future work should explicitly account for this to isolate potential confounds by evaluating the precision of the gesture capture system under varying environments (e.g., lighting conditions or hardware setups) and potential error-handling mechanisms. We can also consider whether other tasks are more suitable for gesture-based input. For instance, gestures as an input modality may be relatively more effective in spatial or exploratory tasks, where natural movement reduces friction or enhances expressiveness. In the same vein, gesture inputs may perform differently in tasks requiring fine motor control versus coarse selections. Future work can aim to develop a theoretical framework mapping task types to suitable modalities to enhance generalizability of these findings.

Mitigating Cognitive Biases. Crowdsourced experiments come with the natural risk of cognitive biases that can negatively affect research outcomes if not considered carefully. Therefore, when designing our experiments we utilized the Cognitive Biases Checklist [14] to assess the presence of any biases in our experimental design. With our focus on comparing input modalities, we acknowledge the potential for an *affect heuristic* like the *familiarity bias*. The ubiquitous presence of `Standard` inputs in today's world means that workers will naturally be more familiar with this input modality. We attempt to control this by providing a tutorial stage at the beginning of the task workflow, exposing workers to the new input modality and allowing them to practice its use. The potential for *disaster neglect* is reduced through an explicit informed consent form that workers read and agree to before performing the task. To avoid the *sunk cost fallacy*, we conducted a pilot study to estimate an appropriate amount of time required to complete the task. The *optimism bias* is mitigated via detailed, clear task descriptions and instructions. Finally, *self-interest bias* exists due to the monetary compensation component of the microtask platform used. To mitigate this bias, we excluded participant submissions from our analysis if they failed two or more attention checks, provided the same answer for all Likert-scale questions, and spent less than a minimum of fifteen minutes to complete the study.

5.2 Reflection on the Practical Utility of Our Work and the Broad Societal Impact

Crowd work is pivotal in advancing artificial intelligence (AI) technology, providing large-scale data annotation, validation, and discovery for maintaining data and model quality. Recently, these advancements have taken the form of generative language or diffusion models such as Mistral [25], ChatGPT [33], or MidJourney.[6] Many of these models are pre-trained on content from the internet. As more generated content becomes available online, this training cycle is at risk of degrading. Model performance likely will not improve if future models are trained on generated output. Moreover, not all information on the internet is acceptable for training AI models. Efforts are therefore made to annotate, filter, and moderate such content before training a model [49]. Typically, this annotation is performed through large-scale crowd work and often using microtask crowdsourcing marketplaces. As argued by Gray and Suri [20], there is human labor behind nearly all the advances in AI these days, particularly in data annotation or curation pipelines. It is therefore important to explore new ways in which crowd worker experiences can be improved, how their health and wellbeing can be supported and safeguarded through shaping better work practices, and how barriers to participation in this paradigm can be lowered. Our work in this paper takes an important stride in this spirit by exploring gestures as an alternative input modality for workers to complete their work and advancing the understanding of how integrating gestures alongside conventional input methods will shape task outcomes and workers' experiences.

Our findings suggest that while there is potential for gestures to serve as an effective input modality, there may be task-specific trade-offs to contend with. Crowd work displays similarities to desk work, which comes with associated risks, commonly related to ergonomics and musculoskeletal disorders [9, 35, 41]. Crowd workers operate in multitasking contexts, taking on tasks in groups, and performing repetitive tasks, which creates an environment that poses risks of stress-related injuries like carpal tunnel syndrome. Dubey et al. [15] suggests that stretching exercises can reduce musculoskeletal pain for workers. Proper investigation and design (i.e., addressing the mapping problem) of gestures as effective inputs can bring physical and mental benefits. Experiments to identify and validate benefits related to gesture inputs for crowd work is a future line of research that the crowd computing and human-computer interaction communities would benefit from pursuing.

In summary, through a between-subjects controlled study, we found that integrating gestures as an input modality with the conventional mouse and keyboard modality in batch sequences did not significantly affect all task outcomes or worker experiences in two distinct types of tasks. Our work presents important insights highlighting the potential of considering gestures as an alternative input modality for crowd work. Our future work will explore how scheduling algorithms can be developed to create intelligent and effective means of allocating input modalities based on worker preferences and task fit. We will develop a *gestures-catalog* that is mapped to typical crowdsourcing task types and a browser plugin to support broader adoption.

[6] https://www.midjourney.com/home

Acknowledgments

We thank all the anonymous workers from the Prolific platform who participated in our study. This work was supported by the TU Delft Design@Scale AI lab, the *ProtectMe* Convergence Flagship project, and the TU Delft AI initiative.

References

[1] Alexander T Adams, Elizabeth L Murnane, Phil Adams, Michael Elfenbein, Pamara F Chang, Shruti Sannon, Geri Gay, and Tanzeem Choudhury. 2018. Keppi: A tangible user interface for self-reporting pain. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[2] Alan Aipe and Ujwal Gadiraju. 2018. Similarhits: Revealing the role of task similarity in microtask crowdsourcing. In *Proceedings of the 29th on Hypertext and Social Media*. 115–122.

[3] 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.
[4] Andy Alorwu, Saiph Savage, Niels van Berkel, Dmitry Ustalov, Alexey Drutsa, Jonas Oppenlaender, Oliver Bates, Danula Hettiachchi, Ujwal Gadiraju, Jorge Goncalves, et al. 2022. Regrow: Reimagining global crowdsourcing for better human-ai collaboration. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.
[5] Agathe Balayn, Natasa Rikalo, Christoph Lofi, Jie Yang, and Alessandro Bozzon. 2022. How can Explainability Methods be Used to Support Bug Identification in Computer Vision Models?. In *CHI Conference on Human Factors in Computing Systems*. 1–16.
[6] Ria Mae Borromeo, Thomas Laurent, Motomichi Toyama, and Sihem Amer-Yahia. 2017. Fairness and transparency in crowdsourcing. In *International Conference on Extending Database Technology (EDBT)*.
[7] Carrie J Cai, Shamsi T Iqbal, and Jaime Teevan. 2016. Chain reactions: The impact of order on microtask chains. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3143–3154.
[8] Alessandro Carfì and Fulvio Mastrogiovanni. 2021. Gesture-Based Human-Machine Interaction: Taxonomy, Problem Definition, and Analysis. *IEEE Transactions on Cybernetics* (2021).
[9] A Chandwani, MK Chauhan, and A Bhatnagar. 2019. Ergonomics assessment of office desk workers working in corporate offices. *IJHSR* 9, 8 (2019), 367–375.
[10] Chun-Wei Chiang, Anna Kasunic, and Saiph Savage. 2018. Crowd coach: Peer coaching for crowd workers' skill growth. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–17.
[11] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)* 51, 1 (2018), 1–40.
[12] 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*. 1–26.
[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] 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.
[15] Neha Dubey, Gaurav Dubey, Himanshu Tripathi, and ZA Naqvi. 2019. Ergonomics for desk job workers-an overview. *Int J Health Sci Res* 9, 7 (2019), 257–266.
[16] 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.
[17] 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.
[18] Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. 2017. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 4 (2017), 1–26.
[19] 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.
[20] Mary L Gray and Siddharth Suri. 2019. *Ghost work: how to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.
[21] 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.
[22] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.
[23] Mokter Hossain and Ilkka Kauranen. 2015. Crowdsourcing: a comprehensive literature review. *Strategic Outsourcing: An International Journal* 8, 1 (2015), 2–22.
[24] Jeff Howe. 2006. The Rise of Crowdsourcing. https://www.wired.com/2006/06/crowds/. Accessed: 17-05-2023.
[25] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).
[26] Yuan Jin, Mark Carman, Ye Zhu, and Yong Xiang. 2020. A technical survey on statistical modelling and design methods for crowdsourcing quality control. *Artificial Intelligence* 287 (2020), 103351.
[27] Ji-Youn Jung, Sihang Qiu, Alessandro Bozzon, and Ujwal Gadiraju. 2022. Great Chain of Agents: The Role of Metaphorical Representation of Agents in Conversational Crowdsourcing. In *CHI Conference on Human Factors in Computing Systems*. 1–22.
[28] Kwangtaek Kim, Joongrock Kim, Jaesung Choi, Junghyun Kim, and Sangyoun Lee. 2015. Depth camera-based 3D hand gesture controls with immersive tactile feedback for natural mid-air gesture interactions. *Sensors* 15, 1 (2015), 1022–1046.
[29] 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.
[30] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2176–2184.
[31] Laura Lascau, Sandy JJ Gould, Duncan P Brumby, and Anna L Cox. 2022. Crowdworkers' Temporal Flexibility is Being Traded for the Convenience of Requesters Through 19 'Invisible Mechanisms' Employed by Crowdworking Platforms: A Comparative Analysis Study of Nine Platforms. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–8.
[32] Vili Lehdonvirta. 2022. *Cloud empires: How digital platforms are overtaking the state and how we can regain control*. MIT Press.
[33] Q Vera Liao, Werner Geyer, Michael Muller, and Yasaman Khazaen. 2020. Conversational interfaces for information search. *Understanding and Improving Information Search: A Cognitive Approach* (2020), 267–287.
[34] Andrew Mao, Ece Kamar, and Eric Horvitz. 2013. Why stop now? predicting worker engagement in online crowdsourcing. In *First AAAI Conference on Human Computation and Crowdsourcing*.
[35] K Murrell. 2012. *Ergonomics: Man in his working environment*. Springer Science & Business Media.
[36] Edward Newell and Derek Ruths. 2016. How one microtask affects another. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3155–3166.
[37] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 188–197.
[38] Zahra Nouri, Nikhil Prakash, Ujwal Gadiraju, and Henning Wachsmuth. 2023. Supporting requesters in writing clear crowdsourcing task descriptions through computational flaw assessment. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 737–749.
[39] 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.
[40] Tacettin Sercan Pekin. 2017. Guidelines and principles for efficient interaction on large touchscreens with collaborative usage. (2017).
[41] Worawan Poochada and Sunisa Chaiklieng. 2015. Ergonomic risk assessment among call center workers. *Procedia Manufacturing* 3 (2015), 4613–4620.
[42] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Improving worker engagement through conversational microtask crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
[43] Francis KH Quek. 1995. Eyes in the interface. *Image and vision computing* 13, 6 (1995), 511–525.
[44] Shruti Sannon and Dan Cosley. 2019. Privacy, Power, and Invisible Labor on Amazon Mechanical Turk. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
[45] S Sheppard. 2017. Managing Data Quality in Observational Citizen Science. (2017).
[46] Aayush Singh, Sebastian Wehkamp, and Ujwal Gadiraju. 2022. SignUpCrowd: Using Sign-Language as an Input Modality for Microtask Crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 184–194.
[47] Carlos Toxtli and Saiph Savage. 2023. Designing AI Tools to Address Power Imbalances in Digital Labor Platforms. In *Torn Many Ways: Politics, Conflict and Emotion in Research*. Springer, 121–137.
[48] 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.
[49] Jennifer Wortman Vaughan. 2018. Making better use of the crowd: How crowdsourcing can advance machine learning research. *Journal of Machine Learning Research* 18, 193 (2018), 1–46.
[50] Tijana Vuletic, Alex Duffy, Laura Hay, Chris McTeague, Gerard Campbell, and Madeleine Grealy. 2019. Systematic literature review of hand gestures used in human computer interaction interfaces. *International Journal of Human-Computer Studies* 129 (2019), 74–94.

[51] Juan Pablo Wachs, Mathias Kölsch, Helman Stern, and Yael Edan. 2011. Vision-based hand-gesture applications. *Commun. ACM* 54, 2 (2011), 60–71.
[52] 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.
[53] Alex C Williams, Gloria Mark, Kristy Milland, Edward Lank, and Edith Law. 2019. The Perpetual Work Life of Crowdworkers: How Tooling Practices Increase Fragmentation in Crowdwork. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–28.
[54] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. 2015. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755* (2015).
[55] Yuxiang Zhao and Qinghua Zhu. 2014. Evaluation on crowdsourcing research: Current status and future direction. *Information systems frontiers* 16 (2014), 417–434.