# Combining Worker Factors for Heterogeneous Crowd Task Assignment

Senuri Wijenayake
The University of Sydney
Sydney, Australia
senuri.wijenayake@sydney.edu.au

Danula Hettiachchi
RMIT University
Melbourne, Australia
danula.hettiachchi@rmit.edu.au

Jorge Goncalves
The University of Melbourne
Melbourne, Australia
jorge.goncalves@unimelb.edu.au

## ABSTRACT

Optimising the assignment of tasks to workers is an effective approach to ensure high quality in crowdsourced data - particularly in heterogeneous micro tasks. However, previous attempts at heterogeneous micro task assignment based on worker characteristics are limited to using cognitive skills, despite literature emphasising that worker performance varies based on other parameters. This study is an initial step towards understanding whether and how multiple parameters such as cognitive skills, mood, personality, alertness, comprehension skill, and social and physical context of workers can be leveraged in tandem to improve worker performance estimations in heterogeneous micro tasks. Our predictive models indicate that these parameters have varying effects on worker performance in the five task types considered – sentiment analysis, classification, transcription, named entity recognition and bounding box. Moreover, we note 0.003 - 0.018 reduction in mean absolute error of predicted worker accuracy across all tasks, when task assignment is based on models that consider all parameters vs. models that only consider workers' cognitive skills. Our findings pave the way for the use of holistic approaches in micro task assignment that effectively quantify worker context.

## CCS CONCEPTS

• **Human-centered computing** → **Computer supported cooperative work**; • **Information systems** → **Crowdsourcing**.

## KEYWORDS

crowdsourcing, task assignment, worker factors, performance

**ACM Reference Format:**
Senuri Wijenayake, Danula Hettiachchi, and Jorge Goncalves. 2023. Combining Worker Factors for Heterogeneous Crowd Task Assignment. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3543507.3583190

## 1 INTRODUCTION

Crowdsourcing provides cheaper, faster and easier access to a massive workforce with diverse capabilities and expertise in comparison to traditional data collection methods. Consequently, crowdsourcing is increasingly been used for Machine Learning research to curate training datasets that feed into different machine learning models that often make critical decisions [70, 81, 82]. Therefore, ensuring quality of crowdsourced data (quality control) - particularly in situations where task requesters have limited visibility of workers' background and skills - has become an interest of many researchers [14, 35, 51, 82].

A recent survey broadly categorised quality control methods as pre-execution (*e.g.*, improving task design and training workers [20, 27]), post-processing (*e.g.*, filtering workers after data collection [60]), and online methods - the latter being particularly effective in heterogeneous micro task environments [35]. Among different online methods that have been recommended for quality control, "task assignment" or dynamically matching workers with micro tasks that are most suitable for them has been extensively researched [28, 34, 37, 39]. However, the task assignment literature predominantly focuses on assessing worker suitability based on their *cognitive ability* - including but not limited to *i.e.*, cognitive flexibility, working memory and inhibition control [28, 34, 37]. While this approach has been successful in improving worker performance in comparison to Expectation Maximisation based (*e.g.*, QASCA [86]) and history-based methods (*e.g.*, 1000 HITs completed with an approval rate of 95% or above [67]), its exclusive and hence limited focus on workers' cognitive ability does not account for other worker factors that can also impact their performance.

Therefore, this study investigates whether and how crowd worker performance (and thereby data quality) in heterogeneous micro tasks can be improved by considering worker factors other than their cognitive ability for task assignment. More specifically, we analyse effects of a curated list of worker factors that can impact worker performance - *i.e.*, their mood [87], personality [47, 48], comprehension skills [57], alertness [26, 31], social context [42], workstation [38], and time of day [38] - together with workers' cognitive ability, on worker performance (task accuracy) in five micro tasks (*i.e.*, Sentiment Analysis, Classification, Transcription, Named Entity Recognition, Bounding Box). Moreover, we compare effects of these factors in low vs. high complexity trials in the five micro tasks considered, to account for potential differences [73].

Our results - based on data collected from 315 crowd workers recruited on Mechanical Turk - indicate that predicting worker performance accounting for the aforementioned worker factors in tandem, rather than exclusively focusing on workers' cognitive ability results in more accurate worker performance estimations (with 0.003 - 0.018 reduction in mean absolute error) in all five micro tasks considered. Moreover, a simulated task assignment


*WWW '23, April 30–May 04, 2023, Austin, TX, USA*

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583190

shows that these improved performance estimations can realistically optimise task assignment in the five micro tasks considered. We discuss implications of our findings on the future of task assignment, particularly as micro crowd tasks are becoming more and more heterogeneous.

## 2 BACKGROUND

### 2.1 Worker Factors for Performance Estimation

Researchers have experimented with different methods such as qualification tests [43], reputation scores [67], previous answers for tasks [49, 86] for worker performance estimation. However, the predominant focus has been on using different *worker factors* for this purpose including (but not limited to) their cognitive ability and other skills, mood, personality, social and physical context [35].

*2.1.1 Cognitive Ability.* Researchers have explored using cognitive ability measurements to estimate worker performance in crowdsourcing contexts to optimise micro task assignment [23, 28, 34, 37]. For example, a study by Goncalves et al. [28] used 8 Factor-referenced Cognitive tests (by ETS) [21] to measure visual and fluency-based cognitive ability of 24 individuals and compared their cognitive ability with task performance in typical micro crowd tasks that appeal to visual (*e.g.*, item recognition) and fluency (*e.g.*, sentiment analysis) skills. Despite being conducted in a lab setting with a limited number of participants, findings of the above study suggest the possibility of reliably measuring cognitive skills of workers, that in turn can be used to optimise task assignment in crowdsourcing environments.

Hettiachchi et al. [34] further investigated the use of cognitive ability for performance estimation. They used 5 standard, fast-paced cognitive tests that are more-suited for the dynamics of crowdsourcing environments, to quantify cognitive ability of workers under three brain functions - Inhibition Control, Cognitive Flexibility, and Working Memory. The authors noted specific correlations between the three executive brain functions and micro crowd tasks considered. For instance, workers who demonstrated higher Inhibition Control (ability to control impulsive responses) performed better than others in sentiment analysis tasks, whereas workers with higher Cognitive Flexibility (ability to switch between mental processes) showed higher performance in transcription tasks. A more recent study implemented a dynamic framework that can recommend or assign heterogeneous micro tasks to crowd workers based on their performance in cognitive ability test [37]. They note that both task assignment and recommendation based on worker performance estimations that account for their cognitive ability can significantly improve worker performance compared to a generic or random task assignment in heterogeneous micro tasks.

*2.1.2 Personality.* Following up from studies that indicate correlations between personality traits of individuals and their work performance in offline environments [41, 65], researchers have investigated how crowd worker personality - captured in terms of the Big-five personality traits *i.e*, Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism [45] - impact their performance in a variety of crowdsourcing tasks [40, 47, 48, 61, 62]. For example, studies note that workers who demonstrate higher Openness, Conscientiousness and Agreeableness tend to perform better in relevance labelling tasks, whereas those who indicate higher scores for Neuroticism tend to display poorer performance [47, 48]. On the other hand, workers' Conscientiousness and Extraversion scores have been seen to positively correlate with performance in translation and transcribing tasks, while Neuroticism scores continue to display a negative correlation with task performance [40, 61]. Moreover, research shows that for tasks that require creativity, workers with higher Openness, Conscientiousness are best-suited [62].

*2.1.3 Mood.* Zhuang and Gadiraju [87] found that workers' self-reported mood - captured using the "Pick-A-Mood" (PAM) scale [16] and categorised as pleasant, unpleasant and neutral - impact their perceived engagement and feeling of accomplishment when completing crowd tasks. More specifically, workers who reported to be in a pleasant mood perceived higher benefits from completing tasks in comparison to workers in an unpleasant mood. Furthermore, a study by Morris et al. [59] shows that temporary priming for positive moods by displaying pleasant pictures (*e.g.*, baby pictures) to crowd workers can improve worker performance (output quality) in idea generation tasks.

*2.1.4 Worker Skills.* Prior work has explored using different worker skills *i.e.*, computer literacy and language literacy to optimise task assignment in crowdsourcing environments [57, 61]. For instance, Mavridis et al. [57] discuss the possibility of using a taxonomy-based skill model to estimate worker performance in computer science related tasks, by comparing the distance between the skills that workers possess to skills required to complete a specific task well. The authors use a 58-item multiple-choice quiz to assess worker skills in this experiment. Another study by Mourelatos and Tzagarakis [61] analysed self-reported data on computer and English literacy of crowd workers to note positive correlations between these skills and worker performance in a transcribing task where workers listened to a music sample and transcribed its lyrics (in English).

*2.1.5 Alertness.* Quality of worker submissions to crowd tasks also depend on their alertness and how attentive they are to the task at hand [50, 56, 75]. Consequently, crowdsourcing experiments that use surveys for data collection often include *attention-check* questions to assess whether workers are paying attention during the task [53, 83]. Typically, these questions are based on the task at hand, are objectively verifiable and can be completed without much effort if attentive. However, worker alertness measured through attention-check questions has only been used as a measure of quality control in crowdsourcing environments, where workers who fail attention check questions are filtered and their work rejected, post-submission [29, 30, 50, 56, 75].

*2.1.6 Worker Context.* Worker context in terms of their *social* situation (alone or with others), *workstation* from where they complete the HIT, and the *time of day* can also impact crowd worker performance. For example, Ikeda and Hoashi [42] observe that workers when required to answer a questionnaire after watching a 3-minute video are less likely to complete the task or spend less time working on the task when surrounded by others, in comparison to when they are alone. The authors further note that workers surrounded by others often displayed lower task accuracy, signifying the effect of

workers' social context on their task performance. Moreover, work by Mao et al. [55] and Chandler and Kapelner [8] indicate that the time of day when workers attempt crowd tasks can also impact their engagement and performance. Additionally, in a study that used crowdsourcing to answer healthcare-related questions, authors note highest engagement from crowd workers in mid-morning to afternoon hours followed by evening hours, in contrast to much lower engagement at night [76].

A recent study that surveyed self-reports of AMT crowd workers on their preference for accepting diverse crowd tasks at different times of the day, found that workers generally prefer morning hours the most - followed by evening, afternoon and night hours - to complete HITs [38]. Furthermore, the same study also analysed if workers' would prefer completing specific tasks over others, depending on their workstation (*i.e.*, a dedicated primary workstation, a temporary workstation, or while commuting). While the authors found no significant effects from workers' workstation on their task acceptance, they note that workers generally preferred to complete HITs from a dedicated workstation. However, this study only looked at workers' preference and did not investigate the impact of these contextual factors on worker performance.

## 3 METHOD

In this study, we aim to determine how different worker factors can affect their performance in a set of typical micro crowd tasks. We used different tests to measure workers' cognitive ability, personality, mood, alertness, and comprehension skills. Further, we used questionnaires to capture worker context and time during which the HIT (*i.e.*, Human Intelligence Task) was completed.

### 3.1 Measuring Worker Factors

*3.1.1 Cognitive Tests.* We measure workers' cognitive ability with respect to three executive brain functions - inhibition control, cognitive flexibility and working memory - using five cognitive tests. *Inhibition control* determines our ability to control impulsive (or automatic) responses and take appropriate action based on reasoning [6], and can be measured using Stroop and Flanker tests. *Working memory* is the "amount of information that can be held in mind and used in the execution of cognitive tasks" [10] which we quantify using N-back and Pointing tests. We use a Task Switching test to measure workers' *cognitive flexibility* which is "the readiness with which one can selectively switch between mental processes to generate appropriate behavioural responses" [11]. These cognitive tests were previously used in a series of studies by Hettiachchi et al. [34, 37] to measure crowdworkers' cognitive ability with respect to the three executive functions of the brain.

**Stroop Test [54, 77]**: The Stroop test requires participants to indicate the font colours of a series of words displayed on the screen using their keyboard. The font colour could be red, blue or green and participants can press the first letter of the relevant font colour (e.g., "R" for red) on their keyboard. During the test, participants encounter three types of trials - incongruent, congruent and unrelated. In *incongruent* trials, a colour name is displayed in a different font colour, as shown in the Stroop test example in Figure 2 (a) in the Appendix. Contrastingly, in *congruent* trials the name of the colour matches the display colour. In *unrelated* trials, non-colour words (*e.g.*, monkey, ship) are displayed in either red, blue or green font colours. We had 18 trials in total, with 6 per each trial type. The Stroop effect expects people to be less accurate and slower in *incongruent* trials when compared with *congruent* trials [77].

**Eriksen's Flanker Test [22]**: During the Flanker test, participants see five arrows on screen as shown in Figure 2 (b). Each arrow could point towards left (<) or right (>). In each trial, participants are instructed to click either the right or the left arrow key on their keyboard, to indicate the direction of the third arrow. We included 16 such trials in the experiment, with an equal number of *congruent* and *incongruent* trials. In *congruent* trials all five arrows point in the same direction (*e.g.* >>>>> or <<<<<), whereas in *incongruent* trials the arrow in the middle points in the opposite direction to others (*e.g.* >><>> or <<><<). The task effect is similar to the Stroop test.

**Task Switching Test [58]**: We used 16 trials - each displaying a letter-number combination in one of the squares of a 2 x 2 grid as shown in Figure 2 (c). Depending on the position of the stimuli in this grid, participants should focus on either the letter or the number. More specifically, in trials that display the letter-number combination on the top two squares, participants only respond to the letter and press "N" if it is a vowel (*e.g.*, A, E, I, O, U) and "Y" if it is not. Conversely, if the stimuli is present in one of the lower boxes, their response is only determined by the number - "N" if the number is even (*e.g.*, 2, 4, 6, 8) and "Y" if it is not. Two trial types are used in this test - *repeating* and *switching* trials (8 occurrences each). *Repeating* trials position the stimuli in top or bottom boxes so that participants respond to the letter or the number in consecutive trials. In other words, they would repeatedly focus on either the letter or the number in both trials. On the other hand, *switching* trials would force participants to shift their focus from the letter to the number or vice versa in consecutive trials.

**N-Back Test [66]**: The N-Back test measures the working memory of individuals by asking them to follow a series of stimuli. We used the 3-back version of the test in this study. In other words, participants are asked to indicate whether or not the letter they see on the screen in each trial is what they saw three trials back. They would press "Y" if it is the same letter and "N" if not. If the participant's answer is correct, the bar underneath the displayed letter turns green, and red if it is incorrect. We measured worker performance in 16 such trials, with three additional trials in the beginning of the test to display the first three stimuli.

**Self-ordered Pointing Test [68]**: Similar to the N-Back test, Pointing test also measures working memory of participants by testing their ability to keep track of a sequence of recent actions. As shown in Figure 2 (e), in each trial participants see 3–12 identical squares randomly distributed on the screen. At any given time, one square contains a reward (indicated by a black star in a green background). Participants are instructed to click one square at a time without repeating, until the square with the reward is found. When a box is clicked, if it contains the reward it will briefly turn green as shown in Figure 2 (e). If not, it will either turn grey if its empty or red if the participant has clicked on a previously opened box. The reward switches to a different square each time it is found and the trial ends when the reward has shifted to all the squares in

the trial. The test had five trials - each with more squares than the previous trial.

We present workers with simple and clear instructions and an example of how to complete the relevant test before they attempt each cognitive test. Moreover, trials in cognitive tests other than the Pointing test are set to expire in 3.5 seconds. Prior work notes setting reasonable time limits for cognitive tests in crowdsourcing contexts can reduce worker distraction during tests [37]. We collect response time (in milliseconds), accuracy (a value between 0–1) for each trial in Stroop, Flanker, Task Switching and N-back. For the Pointing test, we collect the number of errors and the average response time per trial.

*3.1.2 Personality.* As a part of the HIT, participants complete a standardised 10-item Big-five personality inventory (BFI-10) shown in Figure 3 (a). The BFI-10 has been previously used by Kazai et al. [47, 48] to quantify personality of crowdworkers in terms of Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Created by Rammstedt and John [69], this test can capture personality traits of individuals while retaining significant levels of reliability and validity (in comparison to the original 44-item version [45]) - in research settings with time constraints - such as crowdwork.

During the test, participants used a scale of 1–5 (1: Disagree strongly to 5: Agree strongly) to self-report how well ten statements describe their personality (see Figure 3 (a) in the Appendix). We compute a score (between 1–5) for each of the five personality traits using the scoring key provided by Kazai et al. [47, 48].

*3.1.3 Mood.* To measure worker moods, we use "Pick-A-Mood" (PAM) shown in Figure 3 (b) - a simple, intuitive, character-based pictorial scale to enable users to self-report their mood in one click [16]. Our decision to use this scale is motivated by crowdsourcing literature that recognise PAM as an ideal tool to capture worker mood in crowdsourcing contexts, where users have low motivation and little time to report their moods [87]. Previous work also indicates that visual representations of moods used in PAM can be accurately interpreted by people of different nationalities which further attests to its validity [16].

PAM includes a neutral mood ("I") and 8 non-neutral moods. The non-neutral moods can be categorised into two main mood groups [16] - *pleasant* (B: Excited, A: Cheerful, H: Relaxed, G: Calm), and *unpleasant* (C: Tense, D: Irritated, E: Sad, F: Bored). During the HIT participants are instructed to select the letter that corresponds to the pictorial representation of the mood that most closely resembles their current mood (or how they feel in that moment).

*3.1.4 Comprehension.* To test comprehension skills of our participants, we use a reading passage recommended for high school students (grades 9–12), that takes about 5–7 minutes to complete. After reading the passage, participants answer five multiple-choice questions (MCQs) based on its content. We use the proportion of correct answers to the five MCQs as a measure of their comprehension skill (a value between 0–1). The comprehension passage, questions and the answers used for the test are extracted from ReadWorks (www.readworks.org) - a popular online learning platform that offers reading comprehension questions, on diverse topics such as world history, geography, art, etc. ReadWorks articles are often used in literature to test comprehension skills of users in diverse contexts [12, 24, 79].

*3.1.5 Alertness.* We use a Psychomotor Vigilance Task (PVT) to capture worker alertness [26, 31] during the HIT. While the original PVT lasts for about 10 minutes [18], more recent literature shows that user response times on a 90-second version of the PVT, strongly correlates with the original 10-minute version [72]. Shorter versions of the PVT that typically last for 90 seconds to 2 minutes are considered appropriate in situations where the 10-minute version may be impractical [4, 52], such as crowdwork.

This study follows Dingler et al. [19]'s PVT test setup, where user "alertness" is measured in terms of their reaction time to simple visual stimulus (a numerical counter), using a PVT test version that can last for about 90 seconds. During the test participants see a numerical counter appearing on an otherwise blank screen in random time intervals (between 2–10 seconds). As shown in Figure 3 (c), we advise participants to press the "SPACE" bar on their keyboard as soon as the counter appears on screen. When "SPACE" bar is pressed, the counter pause for a few seconds before moving to the next trial. The test includes ten such trials. For each trial, we record participant's response time (in milliseconds) - *i.e.*. the time taken by the participant to press "SPACE" from the moment the counter appears on screen) - as a measure of their alertness.

*3.1.6 Worker Context.* We collect contextual information of workers using a post-task survey. In addition to demographic details such as worker's age, gender and highest level of education, the survey inquires their social context (alone or with others) and the type of workstation used to complete the HIT (primary workstation, temporary workstation, or while commuting). Additionally, we derive the time of day during which the worker completed the HIT (*i.e.*, morning, afternoon, evening or night) based on the start and end times of the HIT that Amazon Mechanical Turk (AMT) automatically records. We included these contextual variables in the survey as prior work shows that they are important factors regarding workers' willingness to accept and complete crowd tasks [38, 42].

## 3.2 Crowdsourcing Tasks

Participants complete five micro tasks during the HIT, namely - Classification, Sentiment Analysis, Transcription, Named Entity Recognition, and Bounding Box. These tasks are meticulously chosen based on prior literature that investigate task assignment and recommendation in crowdsourcing platforms [28, 34, 37]. In addition, a report by Pew Research Center [7] indicates that *image classification* tasks that require workers to identify certain pieces of information in images, tag them (with bounding boxes) or classify images based its information content are the most frequently requested (37%) crowd tasks on AMT. Accordingly, in our work we include Classification and Bounding Box tasks where participants complete similar activities. This report further notes that *transcription* tasks and other *text classification activities* (such as Named Entity Recognition tasks) are the second most frequently requested (26%) tasks on Mechanical Turk [7]. Similarly, our HIT include Transcription, Sentiment Analysis (a form of single-label text classification), and Named Entity Recognition tasks.

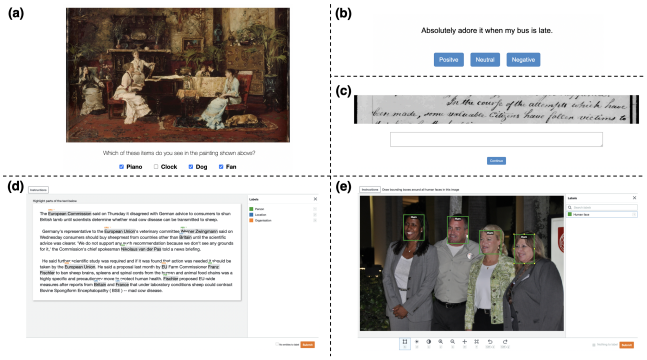


**Figure 1: (a) Classification; (b) Sentiment Analysis; (c) Transcription; (d) Named Entity Recognition; (e) Bounding Box.**

Moreover, we include an equal number of *low* and *high* complexity trials in each task type to investigate the impact of task complexity on how factors in consideration impact worker performance. Task complexity has been reported to impact quality of crowdsourced data in prior work [5, 34, 73]. For each task type, two authors first individually categorised trials into *low* and *high* complexity groups equally. These categorisations were then collated and compared against average trial accuracy of 18 pilot participants who completed these tasks to validate the *low* and *high* complexity trial categories labelled by authors.

*3.2.1 Classification.* In Classification trials, participants are asked to select all items they see in an image, out of a list of four items provided alongside the image as shown in Figure 1 (a). During the HIT, participants complete 16 such trials - each with at least one correct answer. The images used for the test are paintings that represent diverse painting styles from different regions of the world. The chosen set of images and their corresponding answer options have been previously used in crowdsourcing studies [28, 34, 37]. We categorise classification trials as *low* vs. *high* complexity based on the number of items an image contains out of the four answer options provided and how challenging it is to identify them all. For instance, the example provided in Figure 1 (a) is a *high* complexity trial because out of the three correct items participants have to identify - "Piano" and "Dog" are easily spotted, whereas detecting the "Fan" is more challenging.

For each trial, we record participant's response time (in milliseconds) and the number of correct labels they identify. We use the following equation to calculate the accuracy for each trial $t$, with a set of $A$ answers provided by a participant, and a set of $C$ correct answers. Accuracy in each trial is a value between 0 and 1.

$$Accuracy(t, A, C) = max\Big[0, \sum_{a \in A} \frac{1}{|C|} \times \begin{cases} 1, & \text{if } a \in C \\ -1, & \text{otherwise} \end{cases}\Big]$$

*3.2.2 Sentiment Analysis.* Participants complete 16 sentiment analysis trials (extracted from [28, 34, 37]) during the HIT. In each trial, they see a short sentence on the screen (see Figure 1 (b)) and are asked to indicate what sentiment the sentence convey - positive, neutral or negative. We use two types of sentences for this test - straightforward (*low* complexity) and sarcastic (*high* complexity). For example, sentences like "The weather is great today!" convey a clearly positive sentiment, whereas some others like "Absolutely adore it when my bus is late" are sarcastic and hence more challenging to interpret. In addition to the participant's response time (in milliseconds), we record their answer for each trial $a$, to compute trial accuracy (0–1) $t$, when the correct answer is $c$, using the equation below.

$$Accuracy(t, a, c) = \begin{cases} 1, & \text{if } a{=}c \\ 0, & \text{otherwise} \end{cases}$$

*3.2.3 Transcription.* Each Transcription trial presents an image with 2–3 sentences of cursive writing, that participants transcribe to a text box given below the image as shown in Figure 1 (c). This task includes 12 images that have been previously used in crowdsourcing experiments [34, 37]. These images correspond to extracts from The George Washington Papers, representative of inherent individual and period-specific variations in handwriting. Accordingly, we categorise manuscript images with more legible extracts (similar to the example in Figure 1 (c)) as *low* complexity trials and others with less legible extracts as *high* complexity trials. We record the response time and participant response for each trial and compute accuracy for each trial $t$ in terms of the Levenshtein distance (LD) [13] between participant's response string $a$ and the correct answer $c$, using the following equation.

$$Accuracy(t, a, c) = max\Big[0, 1 - \frac{2{\times}LD(a,c)}{string\_length(c)}\Big]$$

*3.2.4 Named Entity Recognition.* The Named Entity Recognition (NER) task includes 10 trials - each displaying brief text passages taken from the publicly available CoNLL-2003 dataset [1]. It includes 1393 English news articles and has been used in prior crowdsourcing experiments [25, 63, 85]. The news articles we chose have 147 words (range: 80–231) and 14 correct tags (range: 11–18) on average. Accordingly, we categorise trials with less than 14 correct tags as *low* complexity and those with more than 14 correct tags as *high* complexity. Moreover, considering that the average reading speed of most adults is around 200 to 250 words per minute, all of these articles can be realistically read in less than a minute.

We used Amazon SageMaker's NER template shown in Figure 1 (d) to integrate this task to the HIT. In each NER trial, we ask participants to read the text carefully, and highlight and tag words or phrases of text that match any one of the following entities - "Person", "Location", and "Organisation". Start and end positions of each tagged word or phrase in the text along with the associated entity are recorded for each tag participants make. We calculate the F1 score for each trial $t$, by comparing a set of participant responses $A$, with the set relevant correct answers $C$ as indicated by the equation below.

$$Accuracy(t, A, C) = 2 \times \Big[\frac{Precision(A,C){\times}Recall(A,C)}{Precision(A,C){+}Recall(A,C)}\Big]$$

*3.2.5 Bounding Box.* In the Bounding Box task, participants are instructed to use a bounding box tool to draw boxes (or rectangles) around human faces in a series of images. Each trial presents an image of people in different social contexts and participants can draw as many rectangles as necessary over each instance of the target (*i.e.* human faces). We use Amazon SageMaker's Bounding Box template shown in Figure 1 (e). The task has 10 trials in total, with half of them displaying images with 2–6 clearly visible human faces

[1] https://paperswithcode.com/dataset/conll-2003

(*low* complexity) and the rest with 10–14 human faces (*high* complexity). These images have been used in previous crowdsourcing experiments [2, 36].

For each Bounding Box trial, we collect participant's response time and position of the rectangles they generate. We then calculate accuracy for each trial $t$, by computing the Intersection Over Union (IOU) score that compares a set of participant responses $A$, with the set relevant correct answers $C$ as indicated by the equation below. IOU score is a metric that is recommended for accuracy computation in Bounding Box tasks [3, 71].

$$Accuracy(t, A, C) = \frac{1}{|C|} \times \left[ \sum_{c \in C} \max(0, IOU(c, a \in A)) \right]$$

## 3.3 Study Deployment

We hosted the experiment on a publicly accessible server with an integrated PostgreSQL database to store worker data. We used psiTurk [32] to integrate the experiment server with AMT seamlessly, meaning that workers were not redirected to an external server. Additionally, several jsPsych plugins [15] and Amazon SageMaker templates [46] were used to create the interfaces used for tests and crowd tasks included in the experiment.

In our study, we integrated all tests and tasks to a single survey and deployed it as a HIT on Amazon Mechanical Turk. We recruited workers who are above 18 years old, fluent in English, and reside in the US. Moreover, eligible workers needed to have completed more than 1000 HITs with an approval rate above 95% - a commonly used qualification criteria in AMT studies [67]. In addition to the above criteria, a pre-qualification survey was used to select workers who have access to a computer with a keyboard (laptop or a desktop computer) to complete the HIT as certain tasks required them to press a key on their keyboard. Eligible workers could preview our task description where we clearly specified that the survey will take approximately 60 minutes to complete (which is the average time taken by 18 pilot participants) and must be completed in a single sitting. Furthermore, workers were provided with the instructions and the requirements of the survey before accepting the task.

Upon accepting the HIT, workers first completed all the tests in a randomised order. Workers then completed the PVT alertness test (Section 3.1.5) immediately before they completed the five crowd tasks described in Section 3.2 - also in a randomised order. The alertness test was positioned in this manner to capture worker's alertness just before they start working on the crowd tasks. The PVT alertness test was not repeated as prior work indicates that repeating the test every two hours is sufficient to continuously assess alertness [78]. Once workers completed the crowd tests, they were presented with the brief post-task survey described in Section 3.1.6 that captured their demographics and contextual details.

The experimental design was approved by the Ethics Committee of our university. We piloted our experimental setup using 18 participants before the survey was deployed to AMT. We then analysed the time spent by pilot participants on each survey item to determine the relevant first quartile (Q1) value. When the experiment was deployed on AMT, these Q1 values were used to determine whether to accept worker submissions or not. More specifically, workers who answered the survey in full and spent sufficient time (above Q1) in 80% of the survey items, received a payment of 15 USD for participation. The payment was decided based on the average time spent by our pilot participants to complete the same survey in a single sitting (60 minutes) and the highest minimum wage in the US [64] at the time of this study (15 USD).

# 4 RESULTS

A total of 354 workers completed the HIT. 315 responses were eligible for further analysis, having spent sufficient time in at least 80% of the survey items. On average workers spent 74 minutes ($SD = 31$) completing the survey. The average time spent by workers completing each test and task included in the survey are provided in the Appendix (Table 1). Additionally, the final sample includes 183 and 132 workers who self-identified themselves as women and men respectively, have completed at least high school (with 84% having completed a Bachelor's degree or a higher qualification), and are between 19–69 years old ($M = 39.6$, $SD = 11.6$).

## 4.1 Outcomes of the Tests Used

*4.1.1 Cognitive Tests.* Worker performance and response time in the five cognitive tests is shown in Figure 4 in the Appendix. On average, accuracy is highest in the Flanker test ($M = 0.76$, $SD = 0.32$) and lowest in the N-Back test ($M = 0.42$, $SD = 0.17$). Additionally, workers have spent the highest and lowest amount of time responding to Task Switching ($M = 1.80$, $SD = 0.58$) and Pointing ($M = 0.95$, $SD = 0.72$) trials.

Moreover, worker accuracy and response times reported for Stroop, Flanker and Task Switching trials establish the presence of corresponding task effects as expected. More specifically, one-sample Wilcoxon signed rank tests show that the difference in accuracy between congruent and incongruent trials is significantly higher than 0 in both Stroop ($V = 7064.5$, $p < 0.001$) and Flanker tests ($V = 12055$, $p < 0.001$). Similarly, the difference in accuracy for repeating and switching trials in the Task Switching test is significantly higher than 0 ($V = 13406.5$, $p = 0.0103$). Additionally, we note that one-sample t-tests show significant differences in response times for congruent and incongruent trials in Stroop ($t(314) = -8.77$, $p < 0.001$) and Flanker tests ($t(314) = -5.53$, $p < 0.001$), as well as for switching and repeating trials in the Task Switching test ($t(314) = -2.43$, $p = 0.015$).

*4.1.2 Personality.* The mean scores reported for Openness ($M = 3.17$, $SD = 0.74$), Conscientiousness ($M = 3.8$, $SD = 0.84$), Extraversion ($M = 3.16$, $SD = 0.93$), Agreeableness and ($M = 3.53$, $SD = 0.90$), and Neuroticism ($M = 2.57$, $SD = 1.00$) are significantly similar to the mean values reported by Kazai et al. [47] in their study analysing the importance of personality for relevance labelling tasks. Accordingly, the higher average scores for Conscientiousness, Agreeableness and the lower score for Neuroticism traits indicate that our worker sample in general tends to perform a thorough job, are trusting and helpful, and emotionally-stable. The borderline mean scores for Openness and Extraversion indicate no particular disposition on these traits in our worker sample.

*4.1.3 Mood, Comprehension, and Alertness.* 80% ($n = 251$) of the workers reported to be in a pleasant mood, with another 15% ($n = 47$) in an unpleasant mood, and the remaining 5% ($n = 17$) in a neutral mood. Our preliminary analysis also indicates comprehension scores in the range of 0–1, with an average score of

0.52 ($SD = 0.25$). This borderline score for the comprehension task indicates average comprehension skill in the recruited worker sample. Additionally, we note an average response time of 0.99 seconds ($SD = 0.86$) for the alertness trials included in the survey, indicating that workers were generally alert as they completed the survey.

*4.1.4 Worker Context.* 77% ($n = 242$) of the workers completed the survey when by themselves, whereas the remaining 23% ($n = 73$) were surrounded by other people as they completed the survey. Additionally, out of the 315 workers in the final sample, 74% completed the survey from a dedicated primary workstation, while another 17% and 9% of the workers were at temporary workstations (*e.g.*, a library or a cafe) and were commuting, respectively. Moreover, 31.5%, 26%, 19% and 23.5% of the workers completed the survey during night, morning, afternoon and evening hours, respectively.

## 4.2 Outcomes of Crowdsourcing Tasks

Worker accuracy (between 0–1) and response time (in seconds) for high and low complexity trials in each of the five crowdsourcing tasks are shown in Figure 5 in the Appendix. One-sample Wilcoxon signed rank tests applied to the difference in average worker accuracy in low vs. high complexity trials indicate that the accuracy differences varied from 0 significantly ($p < 0.001$) in all five task types: Sentiment Analysis - $V = 38148$, Classification - $V = 43202$, Transcription - $V = 41188$, Named Entity Recognition (NER) - $V = 49659$, Bounding Box - $V = 13314$. Similarly, one-sample t-tests show significant differences in response times for high vs. low trials in Sentiment Analysis ($t(314) = 4.85$, $p < 0.001$), Classification ($t(314) = 2.75$, $p = 0.006$), Transcription ($t(314) = 6.23$, $p < 0.001$) and Bounding Box trials ($t(314) = 10.54$, $p < 0.001$). Average response times in high and low complexity NER trials were not significantly different ($t(314) = 1.67$, $p = 0.10$). These results confirm that the low–high trial complexity manipulations used for the five crowdsourcing tasks included in this study is successful.

We note that workers found low complexity NER trials most difficult ($M = 0.28$, $SD = 0.23$), whereas low complexity Classification trials reported the highest mean accuracy ($M = 0.67$, $SD = 0.21$). They took most time completing high complexity transcription trials ($M = 103.65$, $SD = 83.00$), whereas workers were fastest in low complexity Sentiment Analysis trials ($M = 3.76$, $SD = 3.94$).

## 4.3 Predicting Crowd Task Accuracy

We perform model selection using step-wise Generalised Linear Models (GLM) to identify statistically significant effects of the following predictor variables on worker accuracy in the five different crowdsourcing tasks considered in this study. GLMs allow us to identify the effect of a set of predictor variables on an outcome variable (worker accuracy) while following an arbitrary (*i.e.*, possibly non-normal) distribution. Additionally, as we identified significant differences in accuracy for high vs. low complexity trials in all five crowdsourcing tasks, we ran separate GLMs for high vs. low complexity trials in each task type. For each worker we compute:

- **Average accuracy** in Stroop, Flanker, N-Back, Pointing and Task Switching tests (range: 0–1).
- **Average response times** in Stroop, Flanker, N-Back, Pointing and Task Switching tests in seconds.
- **Average test effects** for Stroop, Flanker and Task Switching tests, accuracy (range: -0.5–1) and response time (in seconds).
- **Personality scores** for Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism (range: 1–5).
- **Self-reported mood**: Pleasant, Unpleasant or Neutral.
- **Comprehension score** (range: 0–1).
- **Comprehension response time** in seconds.
- **Average response time in alertness trials** in seconds.
- **Time of day**: Night, Morning, Afternoon, Evening (self-reported data was verified based on the HIT start time).
- **Social context**: By self, With others.
- **Workstation**: Primary, Temporary, Commuting.

All statistically significant predictors ($p < 0.05$) included in the final models with their Mean Absolute Error ($\mathbb{MAE}$), Root Mean Square Error ($\mathbb{RMSE}$) and R-Squared ($\mathbb{R}^2$) values are provided in Table 2 in the Appendix. We report partial eta squared as a measure of the strength of an effect - *i.e.*, 0.01 = small, 0.06 = medium, 0.14 = large - as per Cohen [9] and Winkler et al. [84]. Moreover, final predictors report variance inflation factors well below the often-used threshold of 5 to detect multicollinearity [33].

*4.3.1 Comparing to Previous Work.* As prior work has proposed using cognitive test outcomes for crowd task assignment, we then examine whether additional worker factors explored in our work can provide improved worker performance estimations. For better comparability, similar to Hettiachchi et al. [34], we implement Beta Regression, GLM and Random Forest models with 5-fold cross-validation (100 repeats) and evaluate with $\mathbb{MAE}$, $\mathbb{RMSE}$, and $\mathbb{R}^2$ values. As shown in Figure 6 and Table 3 in the Appendix, predictive models that utilise all the worker factors consistently outperform models that only use cognitive test outcomes.

*4.3.2 Simulated Task Assignment.* Moreover, to investigate whether our task accuracy predictions are useful for practical task assignment, we run a simulated experiment where we select a specific percentage ($K$) of workers for each task, based on their predicted task accuracy. We obtain predicted task accuracy values using our cross-validated Random Forest models (5-folds, 100 repeats). Figure 7 in the Appendix shows the observed (*i.e.* actual) task accuracy of the selected and remaining workers. For example, in bounding box task when $K = 25$, we select 25% of the top performing workers based on predicted task outcomes, resulting in observed mean task outcomes of 0.54 mIOU for selected workers and 0.37 mIOU for the remaining workers. Our simulation demonstrates that consistent worker performance improvements across all five tasks can be obtained through task assignment based on all worker factors.

# 5 DISCUSSION

This study set out to understand if worker performance estimates used for micro crowd task assignment can be improved by considering a combination of worker factors. We find that predictive models that account for different worker factors *i.e.*, personality, mood, alertness, comprehension skill, and social and physical context in tandem with their cognitive ability, outperform models that only account for the latter in estimating worker performance in five different micro crowd tasks. Consequently, these inclusive models can also optimise task assignment despite the heterogeneity of the

micro tasks considered. Therefore, in the context of heterogeneous micro crowd tasks, using a holistic approach that goes beyond workers' cognitive ability can result in more accurate performance estimations for task assignment.

Additionally, except for mood, all worker factors investigated in this study show statistically significant effects on worker performance in at least one task as shown in Table 2. We further note that these worker factors impact worker performance in different micro crowd tasks, at different capacities (as indicated by low–high effect sizes). For instance, workers' comprehension skill that impact their performance in all tasks, shows highest impact in the Named Entity Recognition task that requires them to understand and interpret textual stimuli. We also note that worker context variables (social context, workstation and time of day) are more important in Transcription, Named Entity Recognition and Bounding Box tasks that are generally more time-consuming than the others (see Figure 5). Our findings with regard to the Sentiment Analysis task also imply that effects of certain worker factors like personality can become more evident as task complexity increases.

Therefore, we argue that using a collection of tests to make more holistic performance estimations is crucial to optimise micro task assignment in crowdsourcing platforms. This can be facilitated as an open, test repository framework that holds worker scores in tests they have completed, while also allowing requesters to add new tests as necessary. We emphasise that scores relating to worker factors like cognitive ability, personality and comprehension skills are more durable than some other factors like worker mood, context and alertness that need to be evaluated more often. Hence, these temporal differences should be considered when deciding re-testing requirements. However, as per our results shown in Table 1 (in the Appendix) tests that need to be frequently completed *i.e.*, mood, alertness, and context are less time-consuming than others.

However, for the proposed test repository framework to be effective, a mechanism that can determine task-tests relationships for tasks that are not considered in this study is crucial. Our study investigates and presents task-tests relationships for five most frequently requested micro task types on typical crowdsourcing platforms [7]. Therefore, as an initial step, machine learning models that predict task similarity can be leveraged to expand task-tests relationships we present, to other crowd tasks [1]. As more data on worker factors and their task performance become available organically, these task similarity predictions will naturally improve.

Moreover, we should consider the effort and cost (financial and otherwise) associated with the proposed test repository framework to workers and requesters. Toxtli et al. [80] note that a typical crowd worker already spends approximately 33% of their time on crowdsourcing platforms on unpaid "invisible labour" (*e.g.*, to find appropriate tasks, communicate with requesters, manage payments). Therefore, it is crucial that workers are fairly compensated for the tests they complete, to avoid adding on to "invisible labour". A potential solution would be for the test repository framework to charge a reasonable fee from the requesters who access the test data for task assignment, that can then be used for worker compensation. To motivate the proposed test repository concept from a requester's perspective, we point towards our task assignment results (Figure 7) and prior literature [17, 37, 44] that indicate additional cost of running qualification tests can be recovered by having to recruit fewer, higher quality workers who are better suited for the task. This can also reduce the amount of time and effort requesters spend on post-processing quality control.

Additionally, knowing what tests are necessary to be eligible for a task can allow workers to determine if completing tests is worth the effort [80]. For example, our findings in Table 2 suggest that workers who perform well in the comprehension test are likely to be eligible for all tasks considered in this study (*i.e.* statistically significant, comparatively high effect sizes for comprehension score). Therefore, to encourage workers to complete tests, the framework can indicate potential earning opportunities each test can provide. This can be in the form of other tasks that require the same qualification, which can significantly reduce the "invisible labour" spent by workers searching for tasks that suit their skills [80]. Moreover, having an open, test repository can ensure that workers do not have to repeat the same test to be eligible for similar tasks, unless their scores are no longer applicable. Another approach would be to provide an accurate prediction of how much other workers who completed the same test earned on average, until re-testing is required. Savage et al. [74] has shown that encouraging workers to mimic strategies of high-earning "Super Turkers" - a notion similar to the latter approach - can significantly improve earning opportunities of novice workers.

## 6 CONCLUSION & FUTURE WORK

As crowdsourced data is increasingly being harnessed in life-critical decision making systems, quality control has become more important than ever. Task assignment is an effective quality control mechanism, where optimal task-worker relationships are uncovered based on estimated worker performance and used to assign tasks that are better suited for each individual worker's skill set. In contrast to prior work that focus exclusively on one worker factor, this study proposes estimating worker performance using holistic models that account for diverse worker factors in tandem. Our results assert that inclusive models are more effective for worker performance estimations, particularly as micro crowd tasks are becoming more and more heterogeneous. We discuss implications of our findings for heterogeneous micro task assignment and propose using an open, test repository that records worker factors captured using relevant tests and reuses this data to match workers with tasks that are most suited for their profile.

There are several limitations to our work. While we considered numerous worker factors when investigating task-test relationships across five micro tasks, it is not an exhaustive list of worker factors or crowd tasks. There are other worker factors *i.e.*, behavioural and past performance data, and crowd tasks *i.e.*, audio/video annotation that we did not consider. Furthermore, while the task types we considered were meticulously chosen to be representative of the more frequently requested, heterogeneous micro tasks available on crowdsourcing platforms [7, 34], it is not an exhaustive list of crowd tasks. Additionally, while our results confirm that relationships between certain tests and tasks exist, they do not necessarily mean causation. Therefore, we encourage future work to investigate effects of additional worker factors on more diverse crowd tasks to expand our findings and interpret relationships between test–task performance in depth.

## ACKNOWLEDGMENTS

Danula Hettiachchi is part of the ARC Centre of Excellence for Automated Decision-Making and Society (project number CE200100005), funded by the Australian Government through the Australian Research Council.

## REFERENCES

[1] 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.

[2] Fattoh Al-Qershi, Muhammad Al-Qurishi, Mehmet Sabih Aksoy, Mohammed Faisal, and Mohammed Algabri. 2021. A Time-Series-Based New Behavior Trace Model for Crowd Workers That Ensures Quality Annotation. *Sensors* 21, 15 (2021), 5007.

[3] Saber Mirzaee Bafti, Chee Siang Ang, Md Moinul Hossain, Gianluca Marcelli, Marc Alemany-Fornes, and Anastasios D Tsaousis. 2021. A crowdsourcing semi-automatic image segmentation platform for cell biology. *Computers in Biology and Medicine* 130 (2021), 104204.

[4] Mathias Basner, Daniel Mollicone, and David F Dinges. 2011. Validity and sensitivity of a brief psychomotor vigilance test (PVT-B) to total and partial sleep deprivation. *Acta astronautica* 69, 11-12 (2011), 949–959.

[5] Ria Mae Borromeo, Thomas Laurent, and Motomichi Toyama. 2016. The influence of crowd type and task complexity on crowdsourced work quality. In *Proceedings of the 20th International Database Engineering & Applications Symposium*. 70–76.

[6] Stephanie M Carlson, Louis J Moses, and Casey Breton. 2002. How specific is the relation between executive function and theory of mind? Contributions of inhibitory control and working memory. *Infant and Child Development: An International Journal of Research and Practice* 11, 2 (2002), 73–92.

[7] Pew Research Center. 2016. Research in the Crowdsourcing Age, a Case Study. https://www.pewresearch.org/internet/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/. [Online; accessed 30-August-2020].

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

[9] Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates. 286–288 pages.

[10] Nelson Cowan. 2014. Working memory underpins cognitive development, learning, and education. *Educational psychology review* 26, 2 (2014), 197–223.

[11] Dina R Dajani and Lucina Q Uddin. 2015. Demystifying cognitive flexibility: Implications for clinical and developmental neuroscience. *Trends in neurosciences* 38, 9 (2015), 571–578.

[12] Caitlin Dalton and Joshua Cuevas. 2019. Improving content knowledge in social studies for upper elementary students. *International Journal of Social Sciences & Educational Studies* 5, 3 (2019), 18.

[13] Fred J Damerau. 1964. A technique for computer detection and correction of spelling errors. *Commun. ACM* 7, 3 (1964), 171–176.

[14] 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.

[15] Joshua R De Leeuw. 2015. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods* 47, 1 (2015), 1–12.

[16] 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.

[17] 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.

[18] David F Dinges and John W Powell. 1985. Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior research methods, instruments, & computers* 17, 6 (1985), 652–655.

[19] Tilman Dingler, Albrecht Schmidt, and Tonja Machulla. 2017. Building cognition-aware systems: A mobile toolkit for extracting time-of-day fluctuations of cognitive performance. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–15.

[20] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. 2016. Toward a learning science for complex crowdsourcing tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2623–2634.

[21] Ruth B Ekstrom and Harry Horace Harman. 1976. *Manual for kit of factor-referenced cognitive tests, 1976*. Educational testing service.

[22] Barbara A Eriksen and Charles W Eriksen. 1974. Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics* 16, 1 (1974), 143–149.

[23] Michael Feldman and Abraham Bernstein. 2014. Cognition-based task routing: towards highly-effective task-assignments in crowdsourcing settings. (2014).

[24] Nur Lailatul Fithriyah. 2021. Fostering Students' Positive Attitude Towards Reading Comprehension Through ReadWorks. In *International Seminar on Language, Education, and Culture (ISoLEC 2021)*. Atlantis Press, 236–241.

[25] Hege Fromreide, Dirk Hovy, and Anders Søgaard. 2014. Crowdsourcing and annotating NER for Twitter# drift. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. 2544–2547.

[26] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 1631–1640.

[27] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2017. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM conference on hypertext and social media*. 5–14.

[28] Jorge Goncalves, Michael Feldman, Subingqian Hu, Vassilis Kostakos, and Abraham Bernstein. 2017. Task routing and assignment in crowdsourcing based on cognitive abilities. In *Proceedings of the 26th International Conference on World Wide Web Companion*. 1023–1031.

[29] Jorge Goncalves, Denzil Ferreira, Simo Hosio, Yong Liu, Jakob Rogstadius, Hannu Kukka, and Vassilis Kostakos. 2013. Crowdsourcing on the Spot: Altruistic Use of Public Displays, Feasibility, Performance, and Behaviours. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. 753–762.

[30] Jorge Goncalves, Simo Hosio, Jakob Rogstadius, Evangelos Karapanos, and Vassilis Kostakos. 2015. Motivating participation and improving quality of contribution in ubiquitous crowdsourcing. *Computer Networks* 90 (2015), 34–48.

[31] Anja S Göritz, Kathrin Borchert, and Matthias Hirth. 2021. Using attention testing to select crowdsourced workers and research participants. *Social Science Computer Review* 39, 1 (2021), 84–104.

[32] Todd M Gureckis, Jay Martin, John McDonnell, Alexander S Rich, Doug Markant, Anna Coenen, David Halpern, Jessica B Hamrick, and Patricia Chan. 2016. psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods* 48, 3 (2016), 829–842.

[33] Joseph F Hair, William C Black, Barry J Babin, Rolph E Anderson, and RL Tatham. 2010. *Multivariate Data Analysis*. Pearson, New Jersey, NJ, USA.

[34] Danula Hettiachchi, Niels van Berkel, Simo Hosio, Vassilis Kostakos, and Jorge Goncalves. 2019. Effect of cognitive abilities on crowdsourcing task performance. In *IFIP Conference on Human-Computer Interaction*. Springer, 442–464.

[35] Danula Hettiachchi, Vassilis Kostakos, and Jorge Goncalves. 2022. A Survey on Task Assignment in Crowdsourcing. *ACM Comput. Surv.* 55, 3, Article 49 (2022), 35 pages.

[36] Danula Hettiachchi, Mark Sanderson, Jorge Goncalves, Simo Hosio, Gabriella Kazai, Matthew Lease, Mike Schaekermann, and Emine Yilmaz. 2021. Investigating and Mitigating Biases in Crowdsourced Data. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. 331–334.

[37] Danula Hettiachchi, Niels Van Berkel, Vassilis Kostakos, and Jorge Goncalves. 2020. CrowdCog: A Cognitive skill based system for heterogeneous task assignment and recommendation in crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–22.

[38] Danula Hettiachchi, Senuri Wijenayake, Simo Hosio, Vassilis Kostakos, and Jorge Goncalves. 2020. How context influences cross-device task acceptance in crowd work. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 53–62.

[39] Chien-Ju Ho and Jennifer Vaughan. 2012. Online task assignment in crowdsourcing markets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 26. 45–51.

[40] Kousaku Igawa, Kunihiko Higa, and Tsutomu Takamiya. 2016. An exploratory study on estimating the ability of high skilled crowd workers. In *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, 735–740.

[41] Kousaku Igawa, Kunihiko Higa, and Tsutomu Takamiya. 2020. Utilizing short version big five traits on crowdsouring. *International Journal of Crowd Science* 4, 2 (2020), 117–132.

[42] Kazushi Ikeda and Keiichiro Hoashi. 2017. Crowdsourcing GO: Effect of worker situation on mobile crowdsourcing performance. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 1142–1153.

[43] Panagiotis G Ipeirotis and Evgeniy Gabrilovich. 2014. Quizz: targeted crowdsourcing with a billion (potential) users. In *Proceedings of the 23rd international conference on World wide web*. 143–154.

[44] Ayush Jain, Akash Das Sarma, Aditya Parameswaran, and Jennifer Widom. 2017. Understanding workers, developing effective tasks, and enhancing marketplace dynamics: a study of a large crowdsourcing marketplace. *Proceedings of the VLDB Endowment* 10, 7 (2017), 829–840.

[45] Oliver P John, Sanjay Srivastava, et al. 1999. The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. (1999).

[46] Ameet V Joshi. 2020. Amazon's machine learning toolkit: Sagemaker. In *Machine learning and artificial intelligence*. Springer, 233–243.

[47] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2011. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. 1941–1944.

[48] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2012. The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. 2583–2586.

[49] Asif R Khan and Hector Garcia-Molina. 2017. Crowddqs: Dynamic question selection in crowdsourcing systems. In *Proceedings of the 2017 ACM International Conference on Management of Data*. 1447–1462.

[50] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 453–456.

[51] Matthew Lease. 2011. On quality control and machine learning in crowdsourcing. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

[52] Sylvia Loh, Nicole Lamond, Jill Dorrian, Gregory Roach, and Drew Dawson. 2004. The validity of psychomotor vigilance tasks of less than 10-minute duration. *Behavior Research Methods, Instruments, & Computers* 36, 2 (2004), 339–346.

[53] Henrietta Lyons, Senuri Wijenayake, Tim Miller, and Eduardo Velloso. 2022. What's the Appeal? Perceptions of Review Processes for Algorithmic Decisions. In *CHI Conference on Human Factors in Computing Systems*. 1–15.

[54] Colin M MacLeod. 1991. Half a century of research on the Stroop effect: an integrative review. *Psychological bulletin* 109, 2 (1991), 163.

[55] 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*.

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

[57] Panagiotis Mavridis, David Gross-Amblard, and Zoltán Miklós. 2016. Using hierarchical skills for optimized task assignment in knowledge-intensive crowdsourcing. In *Proceedings of the 25th International Conference on World Wide Web*. 843–853.

[58] Stephen Monsell. 2003. Task switching. *Trends in cognitive sciences* 7, 3 (2003), 134–140.

[59] Robert R Morris, Mira Dontcheva, and Elizabeth M Gerber. 2012. Priming for better performance in microtask crowdsourcing environments. *IEEE Internet Computing* 16, 5 (2012), 13–19.

[60] Yashar Moshfeghi, Alvaro F Huertas-Rosero, and Joemon M Jose. 2016. Identifying careless workers in crowdsourcing platforms: a game theory approach. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 857–860.

[61] Evangelos Mourelatos and Manolis Tzagarakis. 2016. Worker's cognitive abilities and personality traits as predictors of effective task performance in crowdsourcing tasks. In *Proceedings of 5th ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS 2016)*. 112–116.

[62] Pradeep K Murukannaiah, Nirav Ajmeri, and Munindar P Singh. 2016. Acquiring creative requirements from the crowd: Understanding the influences of personality and creative potential in Crowd RE. In *2016 IEEE 24th International Requirements Engineering Conference (RE)*. IEEE, 176–185.

[63] An T Nguyen, Matthew Lease, and Byron C Wallace. 2019. Explainable modeling of annotations in crowdsourcing. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 575–579.

[64] U.S. Department of Labor. 2022. Consolidated Minimum Wage Table. https://www.dol.gov/agencies/whd/mw-consolidated. [Online; accessed 30-August-2020].

[65] Charles A O'Reilly III. 1977. Personality—job fit: Implications for individual attitudes and performance. *Organizational Behavior and Human Performance* 18, 1 (1977), 36–46.

[66] Adrian M Owen, Kathryn M McMillan, Angela R Laird, and Ed Bullmore. 2005. N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human brain mapping* 25, 1 (2005), 46–59.

[67] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior research methods* 46, 4 (2014), 1023–1031.

[68] Michael Petrides, Bessie Alivisatos, Alan C Evans, and Ernst Meyer. 1993. Dissociation of human mid-dorsolateral from posterior dorsolateral frontal cortex in memory processing. *Proceedings of the National Academy of sciences* 90, 3 (1993), 873–877.

[69] Beatrice Rammstedt and Oliver P John. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of research in Personality* 41, 1 (2007), 203–212.

[70] Ashish Rauniyar, Paal Engelstad, Boning Feng, et al. 2016. Crowdsourcing-based disaster management using fog computing in internet of things paradigm. In *2016 IEEE 2nd international conference on collaboration and internet computing (CIC)*. IEEE, 490–494.

[71] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 658–666.

[72] Gregory D Roach, Drew Dawson, and Nicole Lamond. 2006. Can a shorter psychomotor vigilance task be usedas a reasonable substitute for the ten-minute psychomotor vigilance task? *Chronobiology international* 23, 6 (2006), 1379–1387.

[73] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. 2011. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *Proceedings of the international AAAI conference on web and social media*, Vol. 5. 321–328.

[74] Saiph Savage, Chun Wei Chiang, Susumu Saito, Carlos Toxtli, and Jeffrey Bigham. 2020. Becoming the super turker: Increasing wages via a strategy from high earning workers. In *Proceedings of The Web Conference 2020*. 1241–1252.

[75] Kim Bartel Sheehan. 2018. Crowdsourcing research: data collection with Amazon's Mechanical Turk. *Communication Monographs* 85, 1 (2018), 140–156.

[76] Max H Sims, Jeffrey Bigham, Henry Kautz, and Marc W Halterman. 2014. Crowdsourcing medical expertise in near real time. *Journal of Hospital Medicine* 9, 7 (2014), 451–456.

[77] J Ridley Stroop. 1935. Studies of interference in serial verbal reactions. *Journal of experimental psychology* 18, 6 (1935), 643.

[78] Benjamin Tag, Andrew W Vargo, Aman Gupta, George Chernyshov, Kai Kunze, and Tilman Dingler. 2019. Continuous alertness assessments: Using EOG glasses to unobtrusively monitor fatigue levels In-The-Wild. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[79] Crista E Tiboldo et al. 2017. The effect of training in question generation on the development of better questions posed by seventh grade science students. (2017).

[80] 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.

[81] Niels Van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M Kelly, and Vassilis Kostakos. 2019. Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[82] Jennifer Wortman Vaughan. 2017. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *J. Mach. Learn. Res.* 18, 1 (2017), 7026–7071.

[83] Senuri Wijenayake, Danula Hettiachchi, Simo Hosio, Vassilis Kostakos, and Jorge Goncalves. 2020. Effect of conformity on perceived trustworthiness of news in social media. *IEEE Internet Computing* 25, 1 (2020), 12–19.

[84] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376781

[85] Xin Zhang, Guangwei Xu, Yueheng Sun, Meishan Zhang, and Pengjun Xie. 2021. Crowdsourcing Learning as Domain Adaptation: A Case Study on Named Entity Recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5558–5570.

[86] Yudian Zheng, Jiannan Wang, Guoliang Li, Reynold Cheng, and Jianhua Feng. 2015. QASCA: A quality-aware task assignment system for crowdsourcing applications. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*. 1031–1046.

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

# A APPENDICES

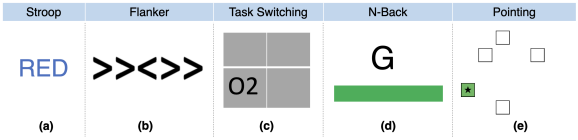


**Figure 2: Examples of cognitive tests used in the study.**

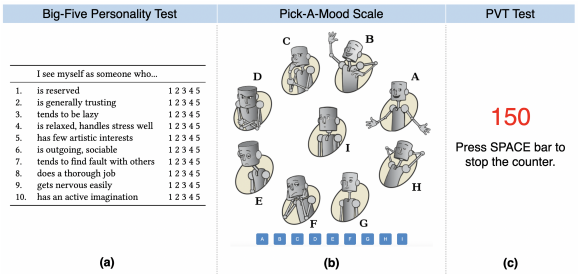


**Figure 3: (a) 10-item BFI [47, 48, 69]; (b)"Pick-A-Mood" scale used to measure worker moods; (c) Interface of the Psychomotor Vigilance Task (PVT) used to capture worker alertness.**

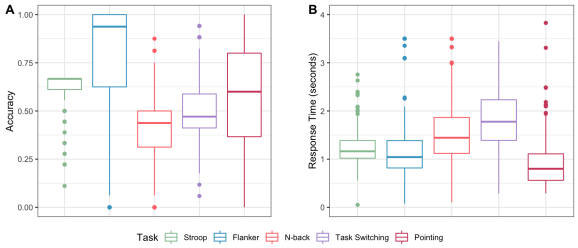


**Figure 4: Accuracy (A), response times (B) for cognitive tests.**

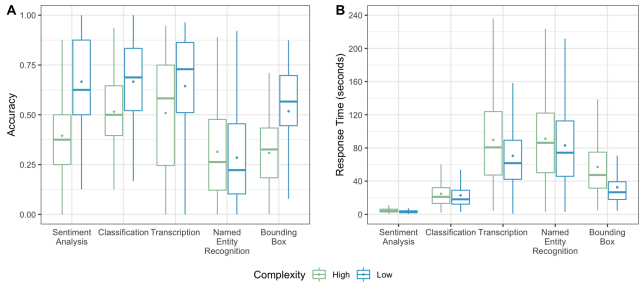


**Figure 5: Accuracy (A), response times (B) for high/low complexity crowdsourcing trials. Mean values indicated as points.**

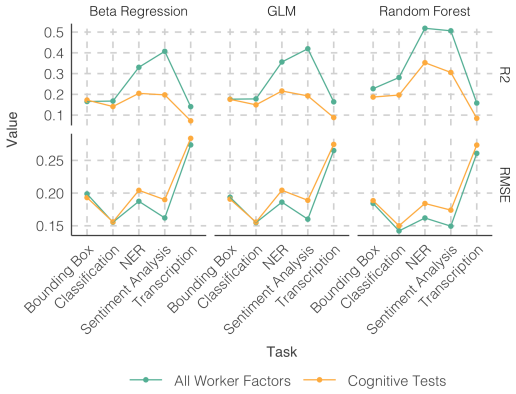


**Figure 6: RMSE and R2 outcomes of Beta Regression, GLM and Random Forest models show that models using all worker factors consistently outperform models that only use cognitive tests.**

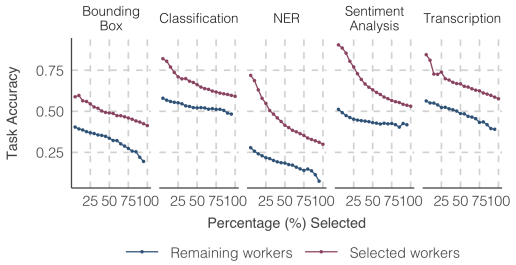


**Figure 7: Observed task accuracy of selected and remaining workers in simulated task assignment. Choosing a subset of workers based on predicted task accuracy can improve overall worker performance across all five tasks.**

| Test/Crowd Task | Average time spent (s) | SD (s) |
|---|---|---|
| **Tests** | | |
| Stroop | 34 | 12 |
| Flanker | 31 | 11 |
| n-Back | 37 | 11 |
| Pointing | 125 | 78 |
| Task Switching | 58 | 121 |
| Comprehension | 443 | 481 |
| Mood | 40 | 137 |
| Personality | 81 | 89 |
| Alertness | 62 | 11 |
| **Crowd Tasks** | | |
| Sentiment Analysis | 74 | 65 |
| Classification | 381 | 242 |
| Transcription | 1061 | 660 |
| Named Entity Recognition | 934 | 536 |
| Bounding Box | 460 | 300 |

**Table 1: Average time (including the time spent on reading the instructions and completing all trials) spent by workers completing different tests and crowdsourcing tasks.**

| Variable | Sentiment Analysis | | Classification | | Transcription | | Named Entity Recognition | | Bounding Box | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Low | High | Low | High | Low | High | Low | High | Low | High |
| Stroop accuracy | – | – | – | – | – | – | $0.02^{*}$ | $0.03^{**}$ | $0.04^{***}$ | $0.04^{***}$ |
| Stroop response time | – | $0.02^{*}$ | – | – | – | – | $0.04^{***}$ | $0.06^{***}$ | $0.03^{**}$ | $0.05^{***}$ |
| Flanker accuracy | – | – | $0.02^{**}$ | – | – | – | $-0.02^{*}$ | – | – | – |
| Flanker response time | $0.01^{*}$ | $0.04^{***}$ | $0.02^{**}$ | – | – | – | – | $0.05^{***}$ | $0.07^{***}$ | $0.06^{***}$ |
| Flanker effect (accuracy) | $0.02^{**}$ | – | – | – | $0.03^{**}$ | $0.03^{**}$ | – | $0.01^{*}$ | – | – |
| N-back response time | – | – | – | – | $0.02^{*}$ | – | – | – | – | – |
| Task switching accuracy | – | – | $0.02^{*}$ | – | – | $0.03^{**}$ | $0.02^{*}$ | $0.03^{**}$ | – | – |
| Task switching response time | $0.05^{***}$ | $0.08^{***}$ | $0.05^{***}$ | $0.02^{**}$ | – | – | $0.02^{*}$ | $0.04^{**}$ | – | – |
| Task switching effect (response time) | – | – | – | – | – | $0.02^{*}$ | – | – | – | – |
| Pointing accuracy | – | – | $0.02^{*}$ | $0.04^{***}$ | – | – | $0.02^{*}$ | – | – | – |
| Comprehension score | $0.11^{***}$ | $0.10^{***}$ | $0.03^{**}$ | $0.04^{***}$ | $0.05^{***}$ | $0.03^{**}$ | $0.18^{***}$ | $0.18^{***}$ | $0.03^{**}$ | $0.05^{***}$ |
| Comprehension response time | $0.01^{*}$ | – | $0.01^{*}$ | – | – | – | – | – | – | – |
| Alertness response time | $0.01^{*}$ | – | $0.02^{*}$ | – | $0.04^{***}$ | $0.03^{**}$ | – | – | – | – |
| Openness | $0.03^{**}$ | $0.07^{***}$ | – | – | – | – | – | – | – | – |
| Conscientiousness | – | $0.04^{***}$ | – | – | – | – | – | – | – | – |
| Extraversion | – | $0.04^{***}$ | $0.01^{*}$ | $0.03^{**}$ | – | – | $0.04^{**}$ | $0.03^{**}$ | – | – |
| Agreeableness | $0.02^{**}$ | – | – | – | – | – | – | – | – | – |
| Neuroticism | – | – | – | – | – | – | $0.01^{*}$ | $0.03^{*}$ | – | – |
| Time of day | – | $0.04^{*}$ | – | – | $0.02^{*}$ | $0.03^{**}$ | $0.01^{*}$ | – | – | $0.02^{**}$ |
| Workstation | – | – | – | – | – | – | $0.02^{*}$ | $0.03^{**}$ | $0.03^{*}$ | $0.03^{**}$ |
| Social context:workstation | $0.04^{***}$ | – | – | – | $0.04^{**}$ | – | – | – | – | – |
| Social context:time of day | – | – | – | $0.04^{**}$ | – | $0.04^{**}$ | – | – | – | – |
| Time of day:workstation | – | – | – | – | – | – | – | – | $0.06^{**}$ | $0.06^{**}$ |
| MAE | 0.14 | 0.15 | 0.13 | 0.12 | 0.18 | 0.21 | 0.13 | 0.14 | 0.16 | 0.12 |
| RMSE | 0.17 | 0.19 | 0.17 | 0.14 | 0.24 | 0.26 | 0.17 | 0.17 | 0.20 | 0.15 |
| $\mathbb{R}^2$ | 0.36 | 0.48 | 0.28 | 0.26 | 0.28 | 0.28 | 0.49 | 0.47 | 0.29 | 0.37 |

**Table 2: Effect sizes (as partial eta square values) of statistically significant predictors in low and high complexity trials for the five crowdsourcing tasks used;** $^{***} = p < 0.001$, $^{**} = p < 0.01$, $^{*} = p < 0.05$

| | | MAE | | RMSE | | R2 | |
|---|---|---|---|---|---|---|---|
| Method | Task | All Tests | Cognitive | All Tests | Cognitive | All Tests | Cognitive |
| Beta Regression | Sentiment Analysis | **0.132** | 0.157 | **0.162** | 0.190 | **0.407** | 0.197 |
| GLM | Sentiment Analysis | **0.130** | 0.156 | **0.160** | 0.189 | **0.420** | 0.193 |
| Random Forest | Sentiment Analysis | **0.121** | 0.139 | **0.149** | 0.174 | **0.506** | 0.306 |
| Beta Regression | Classification | 0.127 | **0.125** | **0.155** | 0.156 | **0.168** | 0.141 |
| GLM | Classification | 0.126 | **0.124** | **0.155** | 0.156 | **0.178** | 0.149 |
| Random Forest | Classification | **0.117** | 0.121 | **0.142** | 0.150 | **0.281** | 0.197 |
| Beta Regression | Transcription | **0.226** | 0.243 | **0.274** | 0.284 | **0.141** | 0.073 |
| GLM | Transcription | **0.209** | 0.222 | **0.265** | 0.274 | **0.164** | 0.088 |
| Random Forest | Transcription | **0.213** | 0.222 | **0.261** | 0.273 | **0.158** | 0.085 |
| Beta Regression | Named Entity Recognition | **0.147** | 0.167 | **0.187** | 0.204 | **0.330** | 0.205 |
| GLM | Named Entity Recognition | **0.146** | 0.167 | **0.186** | 0.204 | **0.356** | 0.216 |
| Random Forest | Named Entity Recognition | **0.132** | 0.147 | **0.162** | 0.184 | **0.518** | 0.352 |
| Beta Regression | Bounding Box | 0.162 | **0.159** | 0.199 | **0.193** | 0.165 | **0.172** |
| GLM | Bounding Box | 0.155 | **0.154** | 0.194 | **0.191** | **0.177** | 0.176 |
| Random Forest | Bounding Box | **0.149** | 0.152 | **0.184** | 0.188 | **0.227** | 0.187 |

**Table 3: MAE, RMSE and R2 values for comparing Cognitive vs. All tests as features in Beta Regression, GLM and Random Forest models with 5-fold cross validation (100 repeats) across all the tasks. Best MAE (lower), RMSE (lower) and R2 (higher) values are given in bold text.**