Effect of Cognitive Abilities on Crowdsourcing Task Performance

Danula Hettiachchi[1], Niels van Berkel[1], Simo Hosio[2], Vassilis Kostakos[1], and Jorge Goncalves[1]

[1] School of Computing and Information Systems, The University of Melbourne, Australia
{danula.hettiachchi, niels.van, vassilis.kostakos, jorge.goncalves}@unimelb.edu.au
[2] Center for Ubiquitous Computing, University of Oulu, Finland
simo.hosio@oulu.fi

Abstract. Matching crowd workers to suitable tasks is highly desirable as it can enhance task performance, reduce the cost for requesters, and increase worker satisfaction. In this paper, we propose a method that considers workers' cognitive ability to predict their suitability for a wide range of crowdsourcing tasks. We measure cognitive ability via fast-paced online cognitive tests with a combined average duration of 6.2 minutes. We then demonstrate that our proposed method can effectively assign or recommend workers to five different popular crowd tasks: Classification, Counting, Proofreading, Sentiment Analysis, and Transcription. Using our approach we demonstrate a significant improvement in the expected overall task accuracy. While previous methods require access to worker history or demographics, our work offers a quick and accurate way to determine which workers are more suitable for which tasks.

Keywords: Crowdsourcing · Cognitive ability · Task performance

1 Introduction

Although crowdsourcing is actively used for a wide variety of both academic and industry tasks, ensuring that the crowd produces data of appropriate quality remains an important challenge. As a result, a wide range of quality assurance mechanisms have been proposed, from straightforward approaches, such as the use of golden standard questions [13] to more complex approaches like monitoring worker activity on crowdsourcing markets [54]. Researchers have also explored ways to predict which workers are likely to perform a task well and facilitate appropriate task assignment [60, 22]. For instance, this can be achieved through the analysis of historical records on completed tasks over a certain period [40, 46]. However, this method is only applicable when such records exist and can be matched to individual workers, which is often not the case. Furthermore, mechanisms that do not rely on the historical performance of workers are better suited in certain scenarios, such as one-time crowdsourcing tasks/campaigns or when

considering new workers of a platform. In these cases, there is no past performance data to predict how well workers would perform on similar or relevant tasks [22].

More robust approaches entail predicting worker performance using different worker attributes, such as age [34], location [34, 56], technical skills [43], and personality [33, 41]. In this paper, we investigate a promising but understudied worker attribute to predict performance in a crowdsourcing setting – cognitive ability. Cognitive ability tests are one of the many methods used by organisations during the recruitment process to identify potential employees with the highest job compatibility. Furthermore, Psychology research has extensively shown that a person's cognitive ability is a good indicator of work performance [55]. In particular, the literature presents three core executive functions of the brain (Inhibition Control, Working Memory, and Cognitive Flexibility) as the basis to describe cognitive ability, which can be measured using appropriate tests [9]. In a crowdsourcing setting, a recent study by Goncalves *et al.* [22] reported promising results regarding the successful prediction of crowd worker performance based on their cognitive skills. However, the completion of the cognitive ability tests (visual and verbal) and crowdsourcing tasks was conducted in a lab study with a limited sample of 24 participants instead of workers from a crowdsourcing platform. Further, the researchers used the Educational Testing Service (ETS) cognitive kit [16], a collection of comprehensive yet complex and time-consuming cognitive tests that are not practical for an online setting. Goncalves *et al.* [22] report that the experiment lasted between 90 to 120 minutes per participant, which would be considered overly long in most online crowdsourcing scenarios.

In this paper we aim to establish a link between the metrics of simple and established online cognitive tests and worker task performance. This link could be used in routing tasks to enhance the efficiency and outcomes of crowd work. As a result, task requesters and crowdsourcing platforms would be able to distinguish the optimum set of workers for a particular crowd task. We conducted an online study on Amazon Mechanical Turk (MTurk)[3] with 102 workers. We asked workers to complete a set of simple and quick (*i.e.,* workers spent on average 6.2 minutes to complete five tests) online cognitive tests (Stroop [42], Flanker [17], N-back [49], Task switching [47], Pointing[50]) that capture the three core executive functions of the brain. This was followed by the completion of typical tasks available in crowdsourcing platforms (Classification, Counting, Proofreading, Sentiment Analysis, Transcription). Our results show a strong relationship between the cognitive ability of crowd workers and their performance in crowdsourcing tasks. We also identify relationships between specific cognitive tests and crowd tasks based on executive functions. Finally, we assign workers to tasks based on their cognitive test scores and demonstrate that our method can significantly improve crowd task accuracy when compared to a baseline generic task assignment.

[3] https://www.mturk.com

2 Related Work

2.1 Human Cognitive Ability and Executive Functions

Human cognitive ability has been extensively studied in Psychology and is often described using executive functions [9]. Executive functions are known to be vital for mental and physical well-being, as well as success in school [3] and at work [2]. The general consensus is that there are three core executive functions: inhibition control, working memory, and cognitive flexibility. These functions form the basis of higher order functions such as reasoning, problem-solving, and planning [9]. *Inhibition control* is the conscious or unconscious restriction of a process or behaviour, especially of impulses or desires. *Working memory* is the ability to hold information in memory and mentally work with it. *Cognitive flexibility* (also known as Switching) is the ability to adapt behaviours in response to changes in the environment and is often associated with creativity [9].

A wide variety of psychological tests such as Stroop [42], Task Switching [47], and N-Back [49] have been developed to assess executive functions. A collection of such tasks is known as a cognitive kit (*e.g.,* Cambridge Neuropsychological Test Automated Battery (CANTAB) [51], Test My Brain [21], The Addenbrooke's Cognitive Examination [45]) and is extensively used in medical and psychological research [9]. Cognitive ability measured from such tests is known to be a good indicator of performance at work, among other predictors such as personality, emotional intelligence, and job experience [55]. This is also well supported by the Person-Job fit theorem which is broadly defined as the compatibility between individuals and jobs [37]. The two aspects of the theory are the suitability of a person for the requirements of a job, and the match between the expectations of a person and the attributes of the job [37]. In theory, any organisation would benefit from optimising their employee selection processes to achieve Person-Job fit, as the literature identifies several positive outcomes such as job performance, satisfaction, and motivation [14].

In a study involving software developers, Chilton *et al.* [4] reported that a misfit between cognitive style and that of the job environment could diminish performance while increasing strain. Similar links between cognitive style and work performance have been established in a number of studies [29, 57]. Although cognitive style or the way individuals think, perceive, and remember information slightly differ from cognitive ability, it correlates with cognitive ability [19]. We also note that several studies have shown that there is no significant relationship between cognitive style and performance at work [53, 38].

In this study we aim to investigate the impact of worker cognitive ability on their task performance in crowdsourcing platforms by measuring cognitive ability using online cognitive tests that capture the three widely established executive functions of the brain.

2.2 Measuring Cognitive Ability Online

Previous work has shown that accurately measuring cognitive ability through online tests is feasible. For instance, Germine *et al.* [21] explored the validity

of using the web for timed, performance-based, and/or stimulus-controlled experiments which are critical for measuring cognitive aptitude online. They reported that web samples do not differ significantly from traditionally recruited or lab-tested samples. Furthermore, participants of their study were anonymous, uncompensated, and unsupervised.

In another example, Crump *et al.* [6] examined the viability of conducting behavioural experiments on crowdsourcing platforms. In a study conducted on MTurk, workers completed tests that are used in cognitive science and cognitive psychology (*e.g.,* Stroop, Flanker, Attentional Blink) with the results being comparable to those collected in laboratory settings. These experiments lasted up to 30 minutes and have characteristics such as multi-trial designs, stimulus presentation, complex instructions, rapid response recording, and requirement of sustained attention of participants. Given these findings and the fact that we based our online cognitive tests on the extensive literature in Psychology on this topic, we anticipate that our online cognitive tests will effectively gauge the cognitive aptitude of crowd workers by testing the three executive functions of the brain.

2.3 Cognitive Ability of Crowdworkers

Eickhoff [15] examined the effect of cognitive biases in crowdsourced relevance labelling tasks and reported that biases could significantly deteriorate the quality of output. A cognitive bias is a systematic error in thinking that affects judgements and decisions. For instance, the framing effect is one such cognitive bias where people respond to a particular option in different ways based on how it is presented. Though cognitive biases differ from cognitive aptitudes, they are closely related and the literature suggests that people with higher cognitive abilities are better at avoiding cognitive biases when making decisions [59].

Alagarai *et al.* [1] investigated different cognitive elements of crowd task design and its effect on performance. They showed that higher task accuracy could be obtained by reducing the demand for visual search and working memory within the task. Previous work by Goncalves *et al.* [22] predicted the accuracy of participants when performing crowd tasks based on cognitive skills measured. However, this experiment was conducted in a laboratory setting, with a small sample, and using the ETS cognitive kit [16], which consists of laborious and time-consuming tests. We aim to investigate this further using straightforward and quick online cognitive tests with a larger sample and explore its applicability for task assignment in crowdsourcing.

2.4 Task Assignment Based on Worker Attributes

Previous work has shown that both demographic and behavioural attributes of workers impact their work quality [33, 34, 41]. In practice, apart from more common attributes such as approval rate, the number of tasks completed, and location, crowd platforms allow requesters to narrow down the worker selection

at a premium price. For example, MTurk allows requesters to select a subset of workers based on worker gender, age, daily internet usage, job, among others.

While there is a strong relationship between crowd worker accuracy and their location in relevance labelling [24, 34] and content analysis [56], studies have confirmed that gender has no significant effect on task accuracy in crowdsourcing [34]. Beyond demographics, personality of the worker is known to affect accuracy. In a study on labelling relevance, Kazai *et al.* [33] segmented crowd workers into five categories based on personality dimensions and reported a significant correlation between personality type and the mean accuracy of the worker. In a subsequent study, Kazai *et al.* [34] also reported that certain personality traits relate to higher task accuracy. Lykourentzou *et al.* [41] examined the effect of personality on the performance of collaborative crowd work on creative tasks and reported that balanced teams containing multiple personalities produce better work in terms of the quality of outcome.

Rzeszotarski and Kittur [54] showed that it is feasible to build predictive models of task performance based on behavioural traces of the user. They introduced a method that analyses the sequence of actions (*e.g.,* mouse movements, scrolling, key-strokes) performed by the user to complete a task, which can be used to measure task accuracy and content quality. Han *et al.* [28] explored annotating the semantic structure of the web using crowdsourcing and reported that most of the behavioural factors of the worker are correlated with the annotation quality. In addition, behaviours of trained professional workers have been successfully used as golden standard to identify those with poor performance [35]. However, behaviour based task performance prediction methods can only be used as post-processing techniques to exclude subpar contributions, which differ from task routing methods. Another approach is to extract the interests of users from social media activity and serve tasks accordingly [11]. We note practical and ethical difficulties in linking worker profiles with social media data.

3 Method

In this study we measured the cognitive ability of crowd workers using five cognitive tests. We then recorded worker performance in five crowdsourcing tasks, and examined if we can utilise cognitive aptitude as an indicator of crowd task performance. We used established cognitive tests to measure the three executive functions of the brain. Table 1 describes the primary executive function measured by each test.

3.1 Cognitive Tests

A description of each cognitive test is provided below.

***Stroop** Test [42].* The classic Stroop test presents two types of trials (incongruent and congruent). As shown in Figure 1, incongruent trials present names of colours (such as "green") displayed in a different colour ("red") whereas congruent trials present names in matching colour. We also included a third trial

Table 1: Cognitive tests and associated executive functions [9]

Cognitive Test	Executive Function
Stroop	Inhibition Control
Flanker	Inhibition Control
Task Switching	Cognitive Flexibility
N-Back	Working Memory
Pointing	Working Memory

type (unrelated) where non-colour words (such as "monkey") appear in either red, green, or blue colour. Participants were asked to press the key corresponding with the first letter of the colour of the word. When asked to focus on the colour of the ink and ignore the meaning of the word (*i.e.,* suppress our prepotent response to words), people are found to be slower and less accurate. This is known as the Stroop effect. Our test contained a total of 18 trials, with a total of 6 trails per type.

Eriksen ***Flanker*** *Test [17].* In each trial crowd workers were presented with a sequence of five arrow symbols (*e.g.,* >>>>>, <<><<) and were asked to pick the centre symbol and press the corresponding arrow key. This task contained 8 congruent (all arrows pointing in the same direction) and 8 incongruent (centre symbol pointing to the opposite direction from the rest) trials. The task effect is similar to the Stroop test.

Task Switching *Test [47].* This test presented a letter and a number in each trial. Depending on whether the pair appears on the upper or lower half of the display, participants were asked to indicate whether the letter is a vowel or consonant, or whether the number is even or odd. The test contained 8 repeating and 8 switching trials.

N-Back *Test [49].* In the N-Back test, crowd workers were presented with a sequence of stimuli. For each stimulus, participants were asked to decide if the current stimulus is the same as the one presented N trials ago, where N can be 1, 2, or 3. We used the 3-back version of this test with each worker completing 16 trials.

Self-ordered ***Pointing*** *Test [50].* In this task, crowd workers were shown 3 to 12 randomly distributed identical squares and were asked to click one box at a time, in any order and without repetition, making sure to click all boxes.

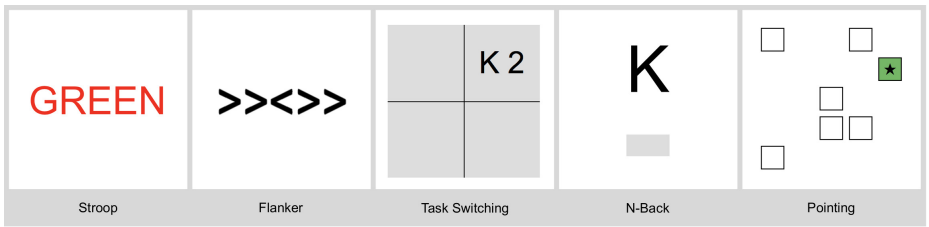

Fig. 1: Screenshots from cognitive tests

Workers received visual feedback after each choice. We tested workers' ability to remember which items they have clicked. The test contained 5 rounds with the total number of squares increasing in each round.

For each test, we specified instructions and included an example prior to the test to ensure workers fully understood the test. Except for the Pointing test, we also configured each trial within the tests to expire after 3.5 seconds. This allowed us to avoid crowd workers pausing the study in the middle of a test and get them to promptly complete each trial. For the Stroop, Flanker, Task Switching, and N-Back tests we recorded accuracy, response time, and trial type (if applicable) for each trial. Based on the trial type, for the Stroop, Flanker, Task Switching tests, test effect was calculated (*e.g.,* Stroop effect in terms of accuracy is the difference in accuracy between congruent and incongruent trials).

3.2 Crowdsourcing Tasks

We used crowdsourcing tasks that are representative of typical tasks available in popular crowdsourcing platforms. Crowd task taxonomy [20] and task availability [10] reported in the literature were also considered. The sentiment analysis and proofreading tasks were adopted from previous work by Goncalves *et al.* [22], and the counting task from Rogstadius *et al.* [52] and Goncalves *et al.* [23, 25]. The transcription and item classification tasks were created specifically for this study. Screenshots from the crowdsourcing tasks are shown in Figure 2 and a description of each task is given below. All tasks had varying complexity as shown in Figure 3 and were presented to participants in random order.

Sentiment Analysis. Crowd workers were asked to identify the sentiment of a sentence (*i.e.,* point of view, opinion). A sentence's sentiment was classified as either 'negative', 'neutral', or 'positive'. The task contained a total of 16 unique sentences. Half of the sentences were straightforward (*e.g.,* "The weather is great today"), while the other half were more challenging due to sentiment ambiguity, context, or sarcasm (*e.g.,*"I'm so pleased road construction woke me up with a bang").

Counting. In this task, workers were presented with an image of a petri dish and asked to count malaria-infected blood cells. Workers were provided with specific instructions on how to differentiate an infected blood cell from an ordinary blood cell. The task contained 8 images that were generated algorithmically con-

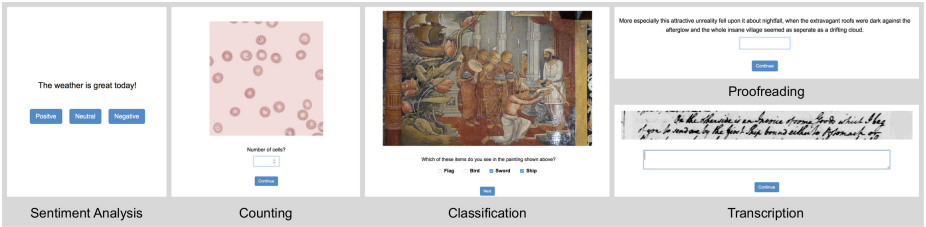

Fig. 2: Screenshots from crowdsourcing tasks

taining varying numbers of infected and ordinary blood cells. Accuracy for each image was determined by $max(0, 1 - \frac{|response - ground\_truth|}{ground\_truth})$.

Item ***Classification***. In this task, crowd workers were presented with 16 paintings (primarily from The Metropolitan Museum of Art[4] and the remaining from Flickr[5], all images licensed for public use) and were asked to identify and mark the items appearing in each painting from a given list of four items. Images represent different painting styles from different countries and contain one or more of the listed items. Certain items could be easily spotted, whereas others were more challenging (*e.g.,* the classification image shown in Figure 2 contains both objects 'Ship' and 'Sword', where the latter is more challenging to locate).

Proofreading. In this task, crowd workers were asked to proofread 12 sentences. Two sentences contained no errors. The remaining sentences contained a single error such as a misspelled word, a grammatical error, or an incorrect word. Workers were asked to type the correct word which should replace the identified erroneous word.

Transcription. Crowdworkers were required to type out a piece of text from a given image. We included 12 images extracted from The George Washington Papers at the Library of Congress [58] in the task. As shown in Figure 3, manuscripts had varying complexity based on the writing style, date, and content. We calculated Levenshtein distance (LD) [7] between the response string and the ground truth and measured accuracy using $max(0, 1 - \frac{2 \times LD}{length(ground\_truth)})$.

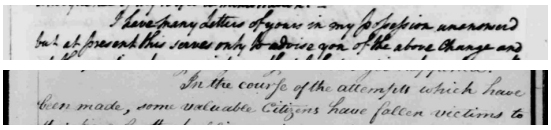

Fig. 3: Transcription tasks of high (top) and low (bottom) complexity.

The cognitive tests were implemented using *jsPsych*, a JavaScript library for online behavioural experiments [39]. Our experiment was integrated with MTurk using *psiTurk* [26], which let us host the experiment on our own server without the need of redirecting users and asking them to submit a completion code.

All tests were encapsulated to a single *Human Intelligent Task* (HIT) and posted to MTurk. When participants accepted the HIT, they were required to electronically sign an informed consent form to start the study. Workers first completed the five cognitive tests, followed by the five crowdsourcing tasks. Both the order of the tests and tasks was randomised. In the last step of the study, participants were requested to provide demographic information (age, gender, and education level). From a pilot study, we estimated that workers would spend around 40 minutes to complete the study. Based on the prevailing federal minimum wage of the United States of $7.25, we payed $5.00 (USD) for each worker

[4] https://www.metmuseum.org/art/collection
[5] https://www.flickr.com

who completed all the tests and tasks. The amount we payed for a worker is comfortably above the average pay one would receive for regular tasks in MTurk [10].

We considered the executive functions associated with each crowdsourcing task during task selection in order to be able to relate them to the different cognitive tests. For example, our counting and classification tasks require sustained attention (Inhibition Control), and demands Working Memory skills while going through the different elements [9]. For the Proofreading task, it is critical to relate to and apply different grammar rules and language patterns (Working Memory and Cognitive Flexibility) [5]. Initially, three of the paper's authors individually identified executive functions linked to each crowdsourcing task based on the literature and their own judgement. The authors then discussed the results, which led to the mapping shown in Table 2.

Table 2: Crowdsourcing tasks and related executive functions

Task	Executive Functions
Classification	Inhibition Control & Working Memory
Counting	Inhibition Control & Working Memory
Proofreading	Working Memory & Cognitive Flexibility
Sentiment Analysis	Cognitive Flexibility & Inhibition Control
Transcription	Cognitive Flexibility & Working Memory

4 Results

A total of 102 workers completed the study (Female 48, Male 54). On average, workers spent 43.6 minutes to complete the study, with 37.0 minutes spent on the crowdsourcing tasks ($SD = 10.7$) and 6.2 minutes on the cognitive tests ($SD = 2.1$). Based on a Pearson Correlation test, we found a significant correlation between the worker scores for the cognitive tests and the mean accuracy for the crowdsourcing tasks ($r = 0.47, p < 0.01$), as shown in Figure 4.

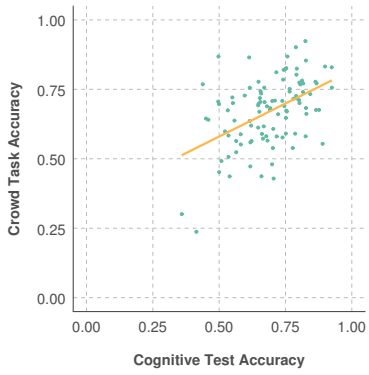

Fig. 4: Accuracy of crowdsourcing tasks vs accuracy of cognitive tests.

4.1 Cognitive Tests

Figure 5 shows worker performance across the five cognitive tests. Workers found the Stroop test to be relatively easier than the rest. In contrast, the mean accuracy of the N-back task is consistently low. Workers are slightly faster in responding to the two tests that measure inhibition control, Stroop and Flanker.

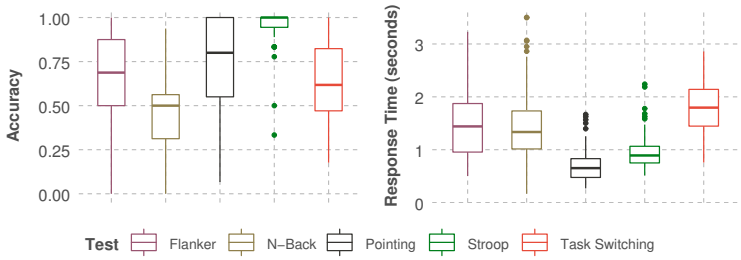

Fig. 5: Accuracy and response time for cognitive tests.

Figure 6 summarises the observed Stroop, Flanker, and Task Switching effects in terms of response time and error rate. As indicated by ANOVA results, for both Stroop and Flanker tests, workers were less error prone ($F(1, 202) = 26.88, p < 0.01$, $F(1, 202) = 8.80, p < 0.01$) and faster ($F(1, 202) = 16.16, p < 0.01$, $F(1, 202) = 5.22, p < 0.05$) when presented with congruent tasks. In the Task Switching test workers were generally faster ($F(1, 202) = 6.78, p < 0.01$) when the same type of task was repeated as opposed to switching from one type to another. This confirms that the effect of the tests was in the expected direction [42, 17, 47].

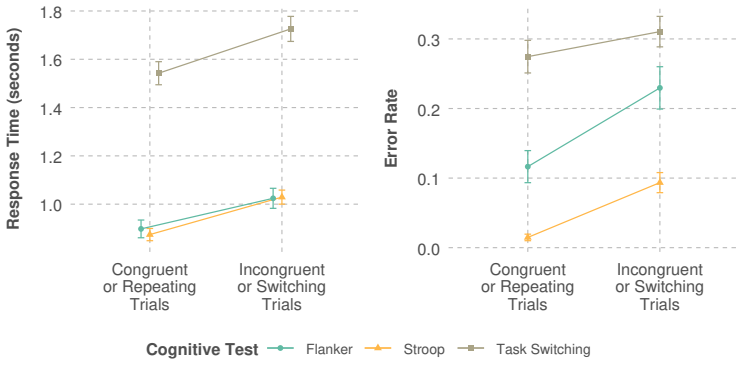

Fig. 6: Stroop, Flanker, and Task Switching effects.

4.2 Crowdsourcing Tasks

Figure 7 shows that workers were generally faster and more accurate in the Sentiment Analysis task as compared to other tasks. Worker accuracy was lowest for the Proofreading task. Figure 8 visualises the accuracy of workers for each sub task of the crowdsourcing tasks (*e.g.,* an individual sentence in the sentiment task). This demonstrates that there is a varying level of complexity within each of our crowdsourcing tasks, an aspect we aimed for in the initial study design. Finally, we do not observe a significant impact of gender, age, or education level on crowd task performance.

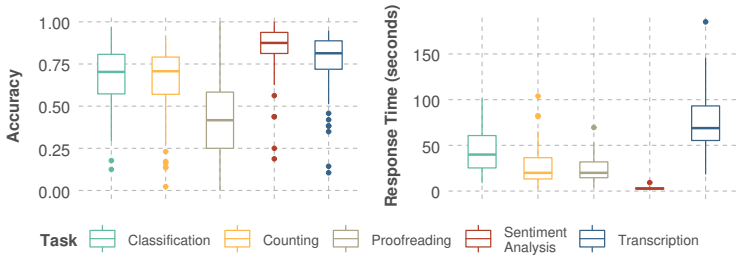

Fig. 7: Accuracy and response time for crowd tasks.

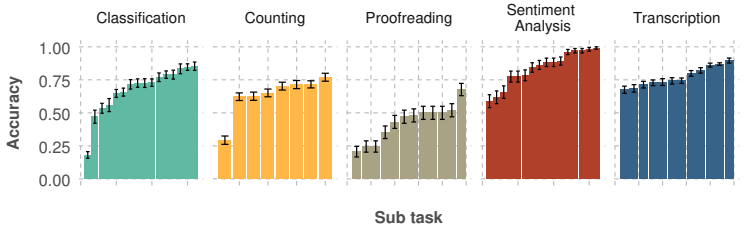

Fig. 8: Accuracy of sub tasks for each crowdsourcing task (Sub tasks are ordered in ascending order of mean accuracy).

4.3 Predicting Crowd Task Accuracy

We used the outcomes of the cognitive tests (*e.g.,* accuracy, response time, Stroop effect) as features to predict the overall accuracy of each worker. Other features include mean response time of instructions and demographic information (age, gender, and education level).

We used Generalised Linear Models, Random Forest, and Beta Regression to predict the overall task accuracy. Mean Absolute Error (MAE), Root Mean

Square Error (RMSE), and R-Squared values for the models with 5-fold cross validation with 10 repeats are shown in Table 3. Inter-correlations were checked prior to constructing the models and the variance inflation factors values of our predictors were below the often-used threshold of 5 to detect multicollinearity [27]. As Beta Regression is optimised for datasets where the output value is in the range (0,1), we had to slightly modify the accuracy values (y) using the equation, $(y * (n - 1) + 0.5)/n$ where n is the number of observations.

Table 3: Results of predictive models (5-fold cross validation with 10 repeats)

Method	MAE	RMSE	R^2
Generalised Linear Model	0.085	0.105	0.320
Random Forest	0.085	0.105	0.303
Beta Regression	0.083	0.105	0.290

We also predicted the accuracy for individual crowdsourcing tasks using the same procedure. Based on the results (MAE, RMSE, and R-Squared values), we selected Random Forest for further investigation and prediction as it produces slightly better results over the other two models in this analysis. Table 4 presents the features that were shown to be the most important based on feature importance scores of Random Forest models and the respective executive functions that those features relate to, as well as the executive functions we hypothesised each crowdsourcing task covers (Table 2).

Table 4: Significant features and related executive functions

Crowd Task	Hypothesis	Significant Features	Imp. Score	Related Executive Functions
Classification	In. Control	Pointing (Accuracy)	4.95	In. Control
	W. Memory	Flanker (Response Time)	3.07	W. Memory
		Stroop (Accuracy)	2.45	
Counting	In. Control	Flanker (Effect Accuracy)	5.57	In. Control
	W. Memory	Pointing (Response Time)	3.72	W. Memory
		Stroop (Accuracy)	3.37	
Proofreading	W. Memory	Task Switching (Accuracy)	7.93	W. Memory
	Cog. Flexibility	Pointing (Accuracy)	5.60	Cog. Flexibility
		Instructions (Response Time)	4.08	
Sen. Analysis	Cog. Flexibility	Stroop (Response Time)	9.68	In. Control
	In. Control	Instructions (Response Time)	6.90	
		Flanker (Effect Accuracy)	5.74	
Transcription	Cog. Flexibility	Task Switching (Accuracy)	3.03	Cog. Flexibility
	W. Memory	Task Switching (Effect Accuracy)	2.98	

In addition, we applied Principal Component Analysis (PCA) separately for both the cognitive test and crowdsourcing task results. PCA can be used to show the distance and relatedness among a population. We visualise this analysis in Figures 9 & 10. These figures, known as variable correlation plots, visualise the relationship between all variables. In Figure 9, we observe that the N-back and Pointing tests are grouped together, implying they are highly correlated. Both tests measure Working Memory. Similarly, Stroop and Flanker tests, which both measure Inhibition Control, are positively correlated as shown in Figure 9. More importantly, Figure 9 confirms that our cognitive test results are in agreement with the literature regarding the measured executive functions (as presented in Table 1). The yellow circle indicates a 100% representation of a variable in the given space. The length of the arrows (close to the edge of the circle) indicates that all variables are well represented in both plots.

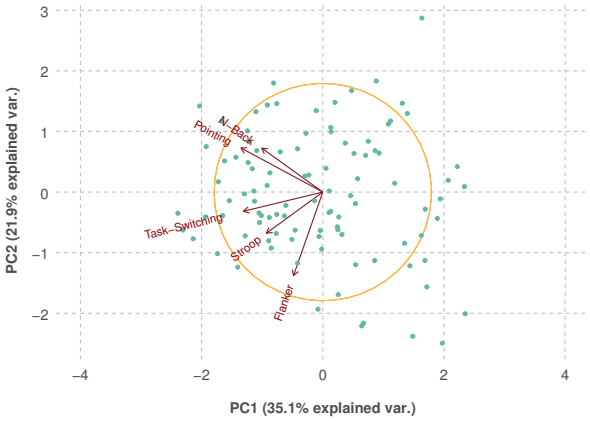

Fig. 9: Principal Component Analysis (PCA) of cognitive tests.

We make two important observations in Figure 10. First, workers are spread throughout the space, which shows the diversity in terms of worker expertise. For example, worker marked as 'W1' in Figure 10 did not perform well on Proofreading and Transcription tasks, but performs above average on Sentiment Analysis and Counting tasks. Our aim is to capture these differences via cognitive tests to facilitate effective task assignment. Second, we identify strong positive correlations among Proofreading and Transcription task pair, and Sentiment Analysis and Counting task pair. This suggests a similarity between tasks in terms of underlying executive functions. According to our findings (Table 4), Cognitive Flexibility is important for both Proofreading and Transcription tasks while Inhibition Control is significant for Sentiment Analysis and Counting tasks.

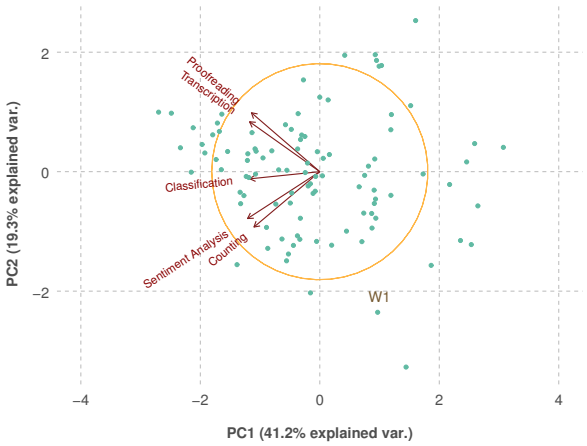

Fig. 10: Principal Component Analysis (PCA) of crowdsourcing tasks.

4.4 Task Assignment Based on Cognitive Skills

Next we developed a strategy to exemplify how cognitive tests can be used for task assignment. To evaluate our strategy, we first select workers for tasks solely based on cognitive test scores, and then compare their task performance as recorded in the study. Here, we transform our prediction from a regression problem to a binary classification problem and focus on predicting if a particular worker should be assigned to a particular crowdsourcing task or not.

For any specific task, we can select a subset of workers from a worker pool in order to maximise the predicted accuracy. For each task, we trained a Random Forest model with 5-fold cross validation using measures from cognitive tests and demographic information as features. Using the models, we predicted the expected accuracy for each worker for each task. Then for each task, based on predicted worker accuracy, we categorised workers into two classes ('Selected' or 'Not Selected'). We used a variable 'Worker Qualification Limit' (L) to determine which portion of workers to consider for assignment. For instance when $L = 40$ for the Classification task, the top 40% of workers in terms of their predicted accuracy in this task are labelled as 'Selected' and the remaining 60% are labelled as 'Not Selected'.

The observed accuracy for workers based on prediction outputs for all five tasks with different L values is shown in Figure 11. For instance, for the Sentiment Analysis task, if we select the top 51 workers out of 102 ($L = 50$) in terms of our predictive model, we observe that those 51 selected workers actually achieve a mean accuracy of 0.88 whereas the 51 unselected workers achieve a mean accuracy of 0.80. The overall mean accuracy for the Sentiment Analysis task for all 102 workers is 0.84 (shown in black horizontal line in Figure 11). Also, we note that for any L value, our assignment method selects a subset of workers whose mean accuracy for the task is better than the mean accuracy of the remaining workers or the mean accuracy of the entire worker pool.

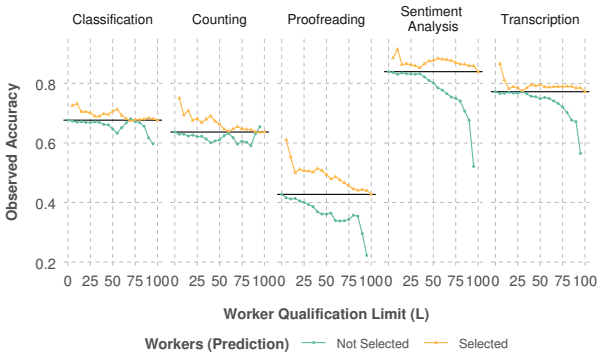

Fig. 11: Accuracy of workers for each task based on output of prediction.

Next we investigated to what extent our method leads to worker discrimination. In other words, does it always favour a handful of skillful workers? We calculate the total number of tasks each worker would be assigned to once we select workers for all five tasks based on our approach. Figure 12 summarises the outcome distribution. If task assignment is carried out based on our model with L as 50, we observer that 11 (10.8%) workers are selected for all five tasks, and 18 (17.6%) workers are not assigned any task. A higher L value (*e.g.,* $L = 75$) assigns more workers to all five tasks. For lower values (*e.g.,* $L = 25$), which represent a more "exclusive" model, we observe that no worker is assigned to all 5 tasks. In other words, at low L values, task routing is so exclusive that there is no single worker in our sample that would meet the expectations for all 5 tasks.

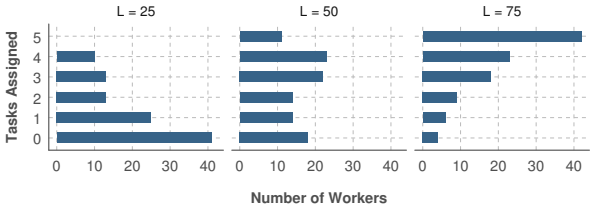

Fig. 12: Number of workers against the total number of tasks assigned to each worker.

5 Discussion

5.1 Using Cognitive Tests to Predict Performance

Apart from cognitive skills, previous work has explored the relationship between crowd task performance and a number of worker attributes, such as age [34], location [34, 56], skills [43], and personality [33, 41]. However, we note a common pitfall in these studies: evaluation is based on a single type of task. For example, Kazai *et al.* [33, 34] only used relevance labelling tasks; Shaw *et al.* [56] used

content analysis questions; Mavridis *et al.* [43] used a set of multiple choice questions on the topic of 'Computer Science'; and Lykourentzou *et al.* [41] explored collaborative advertisement creation. To ensure the applicability of our findings to generic crowd work, our study included five different crowdsourcing tasks.

Previous work by Goncalves *et al.* [22] demonstrated that it is possible to predict the accuracy of crowd workers based on their cognitive skills. While their study used 8 different crowdsourcing tasks to validate their findings, we note three major deficiencies. First, aptitude tests (visual and verbal) as well as the crowdsourcing tasks were conducted in a lab study, using a limited sample of 24 participants that are not representative of the crowd worker population. In contrast, we deployed our entire study on MTurk where 102 actual crowd workers completed the study. Second, compared to the ETS cognitive kit [16] used by Goncalves *et al.* [22], the tests we used to assess the cognitive ability of participants contained fewer trials which were mostly fast-paced. According to the specifications of ETS kit, it takes 44 minutes in total to complete the first part of each cognitive test employed in [22]. In contrast, workers spent on average 6.2 minutes to complete all five of our tests which indicates a significant reduction in required time. Third, unlike ETS tests which are not practical for an online setting (*e.g.,* one task requires paper folding), our online tests can be readily utilised by crowd platforms or task requesters with low effort. Thus, we eliminate any uncertainty associated with the previous study and establish that it is viable to use online cognitive tests to predict crowd task performance.

Furthermore, our prediction model can also be used along with other task routing frameworks. For example, Zheng *et al.* [60] proposed a task assignment system that uses expectation maximisation to populate an *estimated distribution matrix* containing estimated task accuracies. They select optimum tasks to be assigned to a worker based on this matrix. One could easily apply our model based on cognitive skills to predict task accuracy and then generate the estimated distribution matrix.

5.2 Conducting Cognitive Tests Online

We observed Stroop, Flanker, and Task Switching effects that replicate the results of classic Psychology experiments [42, 17, 47]. More importantly, our findings are in line with previous work by Crump *et al.* [6], that demonstrated that these effects could be effectively observed in online experiments. However, the effects we observed indicate a smaller effect size when compared to previous work. One reason for this could be the fact that we used a lower number of trials. For instance, we used 18 trials per worker with 102 workers for Stroop task, whereas the previous work by Crump *et al.* [6] is based on a total of 40 workers, each completing 96 trials.

Furthermore, our study identifies a strong relationship between each crowd task and several cognitive tests, validating our assumption that corresponding executive functions have an extensive impact on the crowd task (see Table 4). Based on this finding, a task requester could either select cognitive tests based on our results or pick executive functions that best explain the nature of the

work and choose matching tests. Alternatively, the requester could implement multiple tests covering all executive functions and then figure out which tests to be used by piloting with a small set of workers. From a crowdsourcing platform perspective, it is more viable to implement a collection of cognitive tests similar to the tests applied in our study, so that the outcomes of such tests can be used to route or recommend a wide variety of tasks to workers.

5.3 Task Assignment

Assigning tasks based on historical performance of workers in crowdsourcing platforms may be impractical for many reasons including anonymity, fluctuations in worker availability [31], or the lack of ground truth data to assess the historical accuracy of workers. On the other hand, using post-processing techniques to reject work could have consequences like workers avoiding the requester in future [44]. Here, we attempt to address these issues by using cognitive tests as predictors of crowd task performance. Our approach for assigning or recommending users, whereby we select a subset of workers who would possibly perform better at each task, can also be seen as a top-N recommendation task [8]. As shown in Figure 12 (for $L = 50$), we observe that tasks are well-distributed amongst workers – despite selecting the best workers for each task. Only 18 workers out of 102 end up not assigned to any task, while 11 workers are selected for all five tasks. This indicates that our proposed model is able to capture different expertise of workers and assign tasks accordingly. Fair task distribution is extremely important when we consider the task assignment problem from the perspective of the crowd workers. In contrast to widely used methods such as approval rate [32], our method does not aim to reward a superior set of workers who are capable in all tasks. Instead, our method focuses on finding the best suited task or tasks for each worker. This will allow workers to complete tasks that are more compatible with their skill set, which has been shown to improve worker satisfaction and reduce the likelihood of task abandonment [36, 30].

Due to budget constraints, crowd task requesters often have to either limit the number of answers expected for each question or reduce the payment for each answer. Both of these actions can reduce output quality [10]. We show that it is possible to obtain higher accuracy by selecting a subset of workers based on cognitive skills (Figure 11), therefore reducing the total number of answers and task cost. In a situation where the requester opts to use cognitive tests as a qualification test, an additional cost would incur for running the cognitive tests. However, typically the number of questions in each task is large enough [31, 10] to recover this initial investment.

5.4 Limitations

We acknowledge several limitations in our study. First, as we wanted to ensure that the cognitive tests took as little time as possible to complete, the total number of trials for each test was kept to a minimum. While measures of cognitive tests would become more accurate and distinct when increasing the number of

trials, the limited number of trials was sufficient for our predictions. Second, human cognitive ability is known to demonstrate subtle variations during the day [12], this is an aspect that we do not account for in our study. Third, similar to any supervised learning method, in the initial stages, our model needs to be trained using data captured from a set of workers performing cognitive tests followed by a set of crowdsourcing tasks similar to those presented in this study.

6 Conclusion and Future Work

In this paper we demonstrate the possibility of using brief online cognitive tests to predict the performance of crowd workers across a range of tasks. We present a study conducted on Amazon Mechanical Turk with 102 workers, where each worker completed a set of cognitive tests followed by a series of crowdsourcing tasks. Through our analysis we highlight the relationships between particular cognitive tests that measure one or more specific executive functions and crowdsourcing task performance.

We show that our proposed method can effectively assign or recommend workers to 5 distinct crowd tasks from a pool of 102 workers with significant improvements to task accuracy while also utilising the majority of the worker pool. Our results also suggest that suitability of a worker for a specific crowdsourcing task could be predicted using the outcome of two or three cognitive tests. Given that each of our cognitive tests could be completed within less than 2 minutes and can be seamlessly integrated with online crowdsourcing platforms, our findings could be readily adopted by researchers, general task requesters, and crowdsourcing platforms.

Further research on the longitudinal impact of the process of measuring cognitive ability would allow us to decide on the optimum frequency with which these tests should be repeated. Cognitive tests should not be repeated too often as it could lead to workers being familiarised with tests. It is known that training obtained in cognitive tests could contribute towards an improvement in metrics of those particular tests but has no impact on other tests or general performance of other tasks [48]. As there are a number of different tests that measure the same executive function [9], one alternative would be to randomly select tests from a pool of tests instead of using identical dedicated tests. In addition, a future study that dynamically routes tasks based on worker cognitive ability and compares the results with other routing methods can further establish the effectiveness of the proposed method in practice.

In our evaluation, we consider assigning workers to tasks one after the other, which will result in repeatedly selecting some workers for multiple tasks. In future work, we intend to explore how we could assign or recommend tasks to workers based on cognitive skills when we have multiple tasks at hand. For this we could either adopt task routing frameworks presented in the literature [60, 40, 18] or propose a novel approach considering additional parameters such as the number of unique questions in each task, the number of answers required for each question, and payment.

References

1. Alagarai Sampath, H., Rajeshuni, R., Indurkhya, B.: Cognitively inspired task design to improve user performance on crowdsourcing platforms. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 3665–3674. CHI '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2556288.2557155
2. Bailey, C.E.: Cognitive accuracy and intelligent executive function in the brain and in business. Annals of the New York Academy of Sciences **1118**, 122–141 (2007). https://doi.org/10.1196/annals.1412.011
3. Borella, E., Carretti, B., Pelegrina, S.: The specific role of inhibition in reading comprehension in good and poor comprehenders. Journal of Learning Disabilities **43**(6), 541–552 (2010). https://doi.org/10.1177/0022219410371676
4. Chilton, M.A., Hardgrave, B.C., Armstrong, D.J.: Person-job cognitive style fit for software developers: The effect on strain and performance. Journal of Management Information Systems **22**(2), 193–226 (2005). https://doi.org/10.1080/07421222.2005.11045849
5. Clair-Thompson, H.L.S., Gathercole, S.E.: Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. The Quarterly Journal of Experimental Psychology **59**(4), 745–759 (2006). https://doi.org/10.1080/17470210500162854
6. Crump, M.J.C., McDonnell, J.V., Gureckis, T.M.: Evaluating amazon's mechanical turk as a tool for experimental behavioral research. PLoS ONE **8**(3), 1–18 (2013). https://doi.org/10.1371/journal.pone.0057410
7. Damerau, F.J.: A technique for computer detection and correction of spelling errors. Communications of the ACM **7**(3), 171–176 (1964)
8. Deshpande, M., Karypis, G.: Item-based top-n recommendation algorithms. ACM Trans. Inf. Syst. **22**(1), 143–177 (Jan 2004). https://doi.org/10.1145/963770.963776
9. Diamond, A.: Executive functions. Annual Review of Psychology **64**(1), 135–168 (2013). https://doi.org/10.1146/annurev-psych-113011-143750
10. Difallah, D.E., Catasta, M., Demartini, G., Ipeirotis, P.G., Cudré-Mauroux, P.: The dynamics of micro-task crowdsourcing: The case of amazon mturk. In: Proceedings of the 24th International Conference on World Wide Web. pp. 238–247. WWW '15, IW3C2, Switzerland (2015). https://doi.org/10.1145/2736277.2741685
11. Difallah, D.E., Demartini, G., Cudré-Mauroux, P.: Pick-a-crowd: Tell me what you like, and i'll tell you what to do. In: Proceedings of the 22nd International Conference on World Wide Web. pp. 367–374. WWW '13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2488388.2488421
12. Dingler, T., Schmidt, A., Machulla, T.: Building cognition-aware systems: A mobile toolkit for extracting time-of-day fluctuations of cognitive performance. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **1**(3) (Sep 2017). https://doi.org/10.1145/3132025
13. Downs, J.S., Holbrook, M.B., Sheng, S., Cranor, L.F.: Are your participants gaming the system?: Screening mechanical turk workers. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 2399–2402. CHI '10, ACM, New York, NY, USA (2010). https://doi.org/10.1145/1753326.1753688
14. Edwards, J.R.: Person-job fit: A conceptual integration, literature review, and methodological critique. John Wiley & Sons, England (1991)

15. Eickhoff, C.: Cognitive biases in crowdsourcing. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. pp. 162–170. WSDM '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3159652.3159654
16. Ekstrom, R.B., Dermen, D., Harman, H.H.: Manual for kit of factor-referenced cognitive tests, vol. 102. Educational Testing Service, Princeton, NJ, USA (1976)
17. Eriksen, B.A., Eriksen, C.W.: Effects of noise letters upon the identification of a target letter in a nonsearch task. Perception & Psychophysics **16**(1), 143–149 (1974)
18. Fan, J., Li, G., Ooi, B.C., Tan, K.l., Feng, J.: icrowd: An adaptive crowdsourcing framework. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. pp. 1015–1030. SIGMOD '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2723372.2750550
19. Federico, P.A., Landis, D.B.: Cognitive styles, abilities, and aptitudes: Are they dependent or independent? Contemporary Educational Psychology **9**(2), 146–161 (1984). https://doi.org/10.1016/0361-476X(84)90016-X
20. Gadiraju, U., Kawase, R., Dietze, S.: A taxonomy of microtasks on the web. In: Proceedings of the 25th ACM Conference on Hypertext and Social Media. pp. 218–223. HT '14, ACM, New York, NY, USA (2014)
21. Germine, L., Nakayama, K., Duchaine, B.C., Chabris, C.F., Chatterjee, G., Wilmer, J.B.: Is the web as good as the lab? comparable performance from web and lab in cognitive/perceptual experiments. Psychonomic Bulletin & Review **19**(5), 847–857 (2012). https://doi.org/10.3758/s13423-012-0296-9
22. Goncalves, J., Feldman, M., Hu, S., Kostakos, V., Bernstein, A.: Task routing and assignment in crowdsourcing based on cognitive abilities. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1023–1031. WWW '17, IW3C2, Switzerland (2017). https://doi.org/10.1145/3041021.3055128
23. Goncalves, J., Ferreira, D., Hosio, S., Liu, Y., Rogstadius, J., Kukka, H., Kostakos, V.: 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. pp. 753–762. UbiComp '13 (2013). https://doi.org/10.1145/2493432.2493481
24. Goncalves, J., Hosio, S., van Berkel, N., Ahmed, F., Kostakos, V.: Crowdpickup: Crowdsourcing task pickup in the wild. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **1**(3), 51:1–51:22 (Sep 2017). https://doi.org/10.1145/3130916
25. Goncalves, J., Hosio, S., Rogstadius, J., Karapanos, E., Kostakos, V.: Motivating participation and improving quality of contribution in ubiquitous crowdsourcing. Comput. Netw. **90**(C), 34–48 (Oct 2015). https://doi.org/10.1016/j.comnet.2015.07.002
26. Gureckis, T.M., Martin, J., McDonnell, J., Rich, A.S., Markant, D., Coenen, A., Halpern, D., Hamrick, J.B., Chan, P.: psiturk: An open-source framework for conducting replicable behavioral experiments online. Behavior Research Methods **48**(3), 829–842 (Sep 2016). https://doi.org/10.3758/s13428-015-0642-8
27. Hair, J., Black, W., Babin, B., Anderson, R.: Multivariate Data Analysis. Prentice-Hall (2010)
28. Han, S., Dai, P., Paritosh, P., Huynh, D.: Crowdsourcing human annotation on web page structure: Infrastructure design and behavior-based quality control. ACM Trans. Intell. Syst. Technol. **7**(4), 56:1–56:25 (Apr 2016)
29. Hoffman, B.J., Woehr, D.J.: A quantitative review of the relationship between person–organization fit and behavioral outcomes. Journal of Vocational Behavior **68**(3), 389–399 (2006). https://doi.org/10.1016/j.jvb.2005.08.003

30. Hosio, S., Goncalves, J., Lehdonvirta, V., Ferreira, D., Kostakos, V.: Situated crowdsourcing using a market model. In: Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology. pp. 55–64. UIST '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2642918.2647362
31. Jain, A., Sarma, A.D., Parameswaran, A., Widom, J.: Understanding workers, developing effective tasks, and enhancing marketplace dynamics: A study of a large crowdsourcing marketplace. Proc. VLDB Endow. **10**(7), 829–840 (2017). https://doi.org/10.14778/3067421.3067431
32. Kazai, G.: In search of quality in crowdsourcing for search engine evaluation. In: Advances in Information Retrieval. pp. 165–176. Springer, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20161-5_17
33. Kazai, G., Kamps, J., Milic-Frayling, N.: Worker types and personality traits in crowdsourcing relevance labels. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. pp. 1941–1944. CIKM '11, ACM, New York, NY, USA (2011). https://doi.org/10.1145/2063576.2063860
34. Kazai, G., Kamps, J., Milic-Frayling, N.: 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. pp. 2583–2586. CIKM '12, ACM, New York, NY, USA (2012). https://doi.org/10.1145/2396761.2398697
35. Kazai, G., Zitouni, I.: Quality management in crowdsourcing using gold judges behavior. In: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. pp. 267–276. WSDM '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2835776.2835835
36. Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., Horton, J.: The future of crowd work. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work. pp. 1301–1318. CSCW '13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2441776.2441923
37. Kristof, A.L.: Person-organization fit: an integrative review of its conceptualizations, measurement, and implications. Personnel Psychology **49**(1), 1–49 (1996)
38. Kristof-Brown, A.L., Zimmerman, R.D., Johnson, E.C.: Consequences of individuals' fit at work: A meta-analysis of person-job, person-organization, person-group, and person-supervisor fit. Personnel Psychology **58**(2), 281–342 (2005)
39. de Leeuw, J.R.: jspsych: A javascript library for creating behavioral experiments in a web browser. Behavior Research Methods **47**(1), 1–12 (2015)
40. Liu, X., Lu, M., Ooi, B.C., Shen, Y., Wu, S., Zhang, M.: Cdas: A crowdsourcing data analytics system. Proc. VLDB Endow. **5**(10), 1040–1051 (Jun 2012)
41. Lykourentzou, I., Antoniou, A., Naudet, Y., Dow, S.P.: Personality matters: Balancing for personality types leads to better outcomes for crowd teams. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. pp. 260–273. CSCW '16, ACM, New York, NY, USA (2016)
42. MacLeod, C.M.: Half a century of research on the stroop effect: An integrative review. Psychological Bulletin **109**(2), 163 (1991)
43. Mavridis, P., Gross-Amblard, D., Miklós, Z.: Using hierarchical skills for optimized task assignment in knowledge-intensive crowdsourcing. In: Proceedings of the 25th International Conference on World Wide Web. pp. 843–853. WWW '16, IW3C2, Switzerland (2016). https://doi.org/10.1145/2872427.2883070
44. McInnis, B., Cosley, D., Nam, C., Leshed, G.: Taking a hit: Designing around rejection, mistrust, risk, and workers' experiences in amazon mechanical turk. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. pp. 2271–2282. CHI '16, ACM, New York, NY, USA (2016)

45. Mioshi, E., Dawson, K., Mitchell, J., Arnold, R., Hodges, J.R.: The addenbrooke's cognitive examination revised (ace-r): a brief cognitive test battery for dementia screening. International Journal of Geriatric Psychiatry **21**(11), 1078–1085 (2006)
46. Mo, K., Zhong, E., Yang, Q.: Cross-task crowdsourcing. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 677–685. KDD '13, ACM, New York, NY, USA (2013)
47. Monsell, S.: Task switching. Trends in Cognitive Sciences **7**(3), 134–140 (2003)
48. Owen, A.M., Hampshire, A., Grahn, J.A., Stenton, R., Dajani, S., Burns, A.S., Howard, R.J., Ballard, C.G.: Putting brain training to the test. Nature **465**, 775 (2010). https://doi.org/10.0.4.14/nature09042
49. Owen, A.M., McMillan, K.M., Laird, A.R., Bullmore, E.: N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. Human Brain Mapping **25**(1), 46–59 (2005). https://doi.org/10.1002/hbm.20131
50. Petrides, M., Alivisatos, B., Evans, A.C., Meyer, E.: Dissociation of human mid-dorsolateral from posterior dorsolateral frontal cortex in memory processing. Proceedings of the National Academy of Sciences **90**(3), 873–877 (1993)
51. Robbins, T.W., James, M., Owen, A.M., Sahakian, B.J., McInnes, L., Rabbitt, P.: Cambridge Neuropsychological Test Automated Battery (CANTAB): A Factor Analytic Study of a Large Sample of Normal Elderly Volunteers. Dementia and Geriatric Cognitive Disorders **5**(5), 266–281 (1994). https://doi.org/10.1159/000106735
52. Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., Vukovic, M.: An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In: Proceedings of the Fifth International AAAI Conference on Web and Social Media. ICWSM, vol. 11, pp. 17–21. AAAI, California, USA (2011)
53. Ruble, T.L., Cosier, R.A.: Effects of cognitive styles and decision setting on performance. Organizational Behavior and Human Decision Processes **46**(2), 283–295 (1990). https://doi.org/10.1016/0749-5978(90)90033-6
54. Rzeszotarski, J.M., Kittur, A.: Instrumenting the crowd: Using implicit behavioral measures to predict task performance. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology. pp. 13–22. UIST '11, ACM, New York, NY, USA (2011). https://doi.org/10.1145/2047196.2047199
55. Schmidt, F.L., Hunter, J.: General mental ability in the world of work: occupational attainment and job performance. Journal of personality and social psychology **86**(1), 162 (2004)
56. Shaw, A.D., Horton, J.J., Chen, D.L.: Designing incentives for inexpert human raters. In: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work. pp. 275–284. CSCW '11, ACM, New York, NY, USA (2011)
57. Verquer, M.L., Beehr, T.A., Wagner, S.H.: A meta-analysis of relations between person–organization fit and work attitudes. Journal of vocational behavior **63**(3), 473–489 (2003). https://doi.org/10.1016/S0001-8791(02)00036-2
58. Washington, G.: George washington papers, series 5, financial papers: Copybook of invoices and letters, 1754-1766 (1766), https://www.loc.gov/item/mgw500003
59. West, R.F., Toplak, M.E., Stanovich, K.E.: Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. Journal of Educational Psychology **100**(4), 930 (2008)
60. Zheng, Y., Wang, J., Li, G., Cheng, R., Feng, J.: Qasca: A quality-aware task assignment system for crowdsourcing applications. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. pp. 1031–1046. SIGMOD '15, ACM, New York, NY, USA (2015)