CrowdCog

A Cognitive Skill based System for Heterogeneous Task Assignment and Recommendation in Crowdsourcing

Hettiachchi, Danula; Van Berkel, Niels; Kostakos, Vassilis; Goncalves, Jorge

Published in:
Proceedings of the ACM on Human-Computer Interaction

DOI (link to publication from Publisher):
10.1145/3415181

Publication date:
2020

Document Version
Accepted author manuscript, peer reviewed version

CrowdCog: A Cognitive Skill based System for Heterogeneous Task Assignment and Recommendation in Crowdsourcing

DANULA HETTIACHCHI, The University of Melbourne, Australia
NIELS VAN BERKEL, Aalborg University, Denmark
VASSILIS KOSTAKOS, The University of Melbourne, Australia
JORGE GONCALVES, The University of Melbourne, Australia

While crowd workers typically complete a variety of tasks in crowdsourcing platforms, there is no widely accepted method to successfully match workers to different types of tasks. Researchers have considered using worker demographics, behavioural traces, and prior task completion records to optimise task assignment. However, optimum task assignment remains a challenging research problem due to limitations of proposed approaches, which in turn can have a significant impact on the future of crowdsourcing. We present 'CrowdCog', an online dynamic system that performs both task assignment and task recommendations, by relying on fast-paced online cognitive tests to estimate worker performance across a variety of tasks. Our work extends prior work that highlights the effect of workers' cognitive ability on crowdsourcing task performance. Our study, deployed on Amazon Mechanical Turk, involved 574 workers and 983 HITs that span across four typical crowd tasks (Classification, Counting, Transcription, and Sentiment Analysis). Our results show that both our assignment method and recommendation method result in a significant performance increase (5% to 20%) as compared to a generic or random task assignment. Our findings pave the way for the use of quick cognitive tests to provide robust recommendations and assignments to crowd workers.

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

Additional Key Words and Phrases: crowdsourcing, dynamic task assignment, cognitive abilities

ACM Reference Format:
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. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 110 (October 2020), 22 pages. https://doi.org/10.1145/3415181

1 INTRODUCTION

The availability of an extensive pool of workers willing and able to complete a high number of tasks has led to crowdsourcing platforms being widely used for data collection efforts by researchers from many different scientific disciplines (*e.g.,* Psychology, Astronomy, Computer Science, Medicine) and by organisations. With the increased use of crowdsourced data in critical applications, researchers have extensively explored approaches to improve the quality of gathered data [12]. While basic approaches such as gold standard questions and qualification tests [18] are commonly used, they

Authors' addresses: Danula Hettiachchi, danula.hettiachchi@unimelb.edu.au, The University of Melbourne, Melbourne, VIC, Australia; Niels van Berkel, nielsvanberkel@cs.aau.dk, Aalborg University, Aalborg, Denmark; Vassilis Kostakos, vassilis.kostakos@unimelb.edu.au, The University of Melbourne, Melbourne, VIC, Australia; Jorge Goncalves, jorge.goncalves@unimelb.edu.au, The University of Melbourne, Melbourne, VIC, Australia.

2573-0142/2020/10-ART110 $15.00
https://doi.org/10.1145/3415181

have inherent limitations which hinder their applicability. For instance, both of the aforementioned methods are task specific and a requester needs to curate the questions or the tests for each task. In addition, crowd tasks often lack ground truth information which makes the creation of gold standard questions challenging. More robust approaches include observing the historic or recent performance of the worker and subsequently estimating the worker's task performance prior to assigning the task [16, 24, 44, 60, 67]. However, in practice, these approaches are ineffective when there is limited or no prior task completion records available. This is particularly problematic when considering the influx of new crowd workers on a platform or one-time crowdsourcing tasks/campaigns. Hence, in our work, we seek to develop a task assignment method which is not based on a worker's prior records and which can be applied across a variety of crowdsourcing task types.

As a single crowdsourcing task is often organised as a collection of sub-tasks or questions of the same task (*e.g.,* translating 50 sentences), task assignment can be further dismantled into two steps: initial task assignment and subsequent question assignment. While task assignment aims to match workers with different types of tasks, question assignment focuses on selecting questions for the worker. The literature shows that crowdsourcing tasks vastly differ in terms of their complexity, required skills, expected time commitment, and allocated payment [15, 26]. Thus, task selection becomes increasingly relevant as the number of tasks and workers available on a crowdsourcing marketplace increases. On the other hand, prior work shows workers often struggle to find compatible or desirable tasks on marketplaces [9]. To match workers with suitable tasks, we investigate the task assignment problem for heterogeneous tasks and use the cognitive skills of crowd workers to predict task performance. Apart from cognitive skills, researchers have also studied the effect of many different worker characteristics like location [43, 63], age [43], personality [42, 49], mood [68] and technical skills [51] on crowd task performance. However, using a worker's cognitive ability for task assignment has a number of benefits over these approaches, such as being straightforward to measure [28], difficult to fabricate [14], and applicable to many task types [29, 36].

Goncalves et al. [29] first showed that cognitive ability can be a good predictor of crowd tasks performance. In addition, recent work by Hettiachchi et al. [36] on Amazon Mechanical Turk (MTurk) shows that online cognitive test performance is correlated with crowdsourcing task performance. Further Hettiachchi et al. [36] propose a model that uses the executive functions of the brain [14] to explain the relationship between cognitive tests and crowdsourcing tasks. However, in their study, workers completed all the cognitive tests and crowdsourcing tasks in a single task unit (*i.e.,* a HIT or Human Intelligent Task in MTurk) which lasts for more than 40 minutes on average. The study does not involve dynamic task routing, but instead conducts an offline analysis of the results. Further, they do not present a system or a framework that demonstrates how cognitive tests can be used to assign tasks or compare the results with any existing task assignment methods. In contrast to the prior work, our experiment replicates typical crowd work conditions where workers have the flexibility to decide on the number of questions they wish to complete and questions are organised into HITs that can be completed in a short time period.

In this paper, we present 'CrowdCog', a real-time online task routing framework that uses cognitive test scores to assign or recommend crowdsourcing tasks to workers. We deploy our study on MTurk and match four different crowdsourcing tasks to workers using the results of five cognitive tests. We show that using our proposed task assignment method, workers are significantly more accurate when compared to a baseline generic task assignment strategy. We also show that workers perform better when following our recommendations instead of selecting the tasks to complete on their own. Further, we compare the performance of our method to a state-of-the-art question assignment method [67] and a standard qualification that uses workers' task completion records. We achieve either similar or better task accuracy through our task routing method that does not use historical data or involve any question selection within the task.

2 RELATED WORK

2.1 Task Assignment

The literature presents a number of task assignment algorithms or frameworks that can be integrated with or used in place of existing crowdsourcing platforms. They consider a variety of different quality metrics (*e.g.,* accuracy, task completion time) and implement one or more quality improvement techniques (*e.g.,* gold standard questions [18], removing erroneous workers [44]) to enhance these quality metrics. The primary motivation behind each assignment method can also be divergent. For example, some methods aim to maximise the quality of the output (*e.g.,* [23, 61, 67]) while other methods attempt to reduce the cost by achieving a reasonable accuracy with a minimum number of workers (*e.g.,* [44]). Task assignment can be further classified into either *heterogeneous task assignment* and *question assignment.*

2.1.1 Task Assignment with Heterogeneous Tasks. As crowdsourcing platforms contain a variety of tasks (*e.g.,* sentiment analysis, classification, transcription), heterogeneous task assignment focuses on matching different task types with workers. Heterogeneous task assignment can be particularly useful in cases where 'expert' workers must be allocated for more difficult tasks [38].

There is limited prior work on heterogeneous task assignment in crowdsourcing. Ho and Vaughan [38] propose a method based on the online primal-dual framework, which has been utilised for different online optimisation problems. In the study, researchers use three types of ellipse classification tasks to account for different expertise levels and use a translation task to simulate different skills. However, their approach assumes that the requester can immediately evaluate the quality of completed work. This vastly limits the applicability of their approach in a real-world crowdsourcing problem. Ho et al. [37] further investigate heterogeneous task assignment in classification tasks with binary labels. However, for the assignment, they use gold standard questions of each task type to estimate the accuracy of the workers.

Assadi et al. [4] studied the task assignment from the requester perspective. They propose an online algorithm that can be used by a requester to maximise the number of tasks allocated with a fixed budget. In a different approach for task assignment, Mo et al. [52] apply a hierarchical Bayesian transfer learning model. They use the historical performance of workers in similar or different type of tasks to estimate the accuracy for the new tasks. Their experiment with a real-world dataset shows the effectiveness of the proposed approach when transferring knowledge from related but different crowd tasks (*e.g.,* questions on sports vs makeup and cooking). However, their real-world evaluation is limited to a single scenario with one source task and one target task. Difallah et al. [16] also propose a system where tasks are allocated based on worker profile data such as interested topics captured from a social media network. The general applicability of this method raises numerous practical and ethical considerations.

While a number of studies have investigated the online task assignment problem, many of them have evaluated only using synthetic data (*e.g.,* [4, 37]). Our study involves a large number of crowd workers and replicates the conditions of typical crowdsourcing platforms.

2.1.2 Question Assignment. Unlike heterogeneous task assignment, the general online task assignment problem has been widely studied in the context of 'question assignment'. In question assignment, which is also often referred to as 'task assignment' (*e.g.,* [44, 67]), the aim is to find the most suitable set of questions from the same task for a given worker.

Zheng et al. [67] propose a task assignment framework, 'QASCA' that uses expectation maximisation on either accuracy or F-score. They experimented on MTurk with five task types including three variants of sentiment labelling of tweets, entity resolution using product descriptions, and selecting which was published earlier from two given films. The method is primarily proposed for multiple

choice questions with a single correct label. QASCA is shown to outperform several other methods including CDAS [48], AskIt! [7], MaxMargin (selecting questions with the highest expected marginal improvement) and ExpLoss (selecting questions based on the expected loss).

'CrowdDQS' is a dynamic task routing mechanism which examines voting patterns and selectively assigns gold standard questions (explicitly verifiable questions) to workers with the aim of identifying and removing workers with poor performance in real-time [44]. The proposed system, which integrates seamlessly with Mechanical Turk, was shown to reduce the number of votes required to accurately answer questions when compared to a round-robin assignment with majority voting. According to the study results, even though CrowdDQS is better than the Expectation Maximisation-based QASCA at worker accuracy estimation, the task accuracy gain is similar.

Fan et al. [23] introduced another dynamic framework named 'iCrowd' that uses a graph-based estimation model to assign tasks to workers with a higher chance of accurately completing the task. They also consider the task similarity when estimating worker accuracy. In another example, Saberi et al. [61] propose a statistical quality control framework (OSQC) for multi-label classification tasks which monitors the performance of workers and removes the workers with high error estimates at the end of processing each batch of tasks. They propose a novel method to estimate the worker accuracy which uses gold standard questions and a plurality answer agreement mechanism. We note that in their evaluation with crowd workers on Amazon's Mechanical Turk, they simulate the past error rates of workers who completed the task, by using a standard normal distribution.

While the literature suggests that these frameworks can produce positive results, their applications are limited for several reasons, such as the fact that these methods have been developed for specific types of crowd work (*e.g.,* [61]) and implemented or tested with a specific crowdsourcing platform (*e.g.,* [44, 61, 67]). One other limitation with regard to benchmarking different methods is the lack of an established crowdsourcing task dataset that spans into different types of crowd tasks.

2.2 Effect of Worker Attributes

When looking at task or question assignment from the workers' perspective, many other worker attributes have been shown to have an impact on crowd task performance. For instance, personality type of the worker is known to be related to the accuracy in relevance labelling tasks [42, 43] and when working in groups [49]. Location of the worker has a significant impact on the task accuracy in content analysis [63] and in relevance labelling [31, 32, 43]. While these studies do not attempt to match workers to tasks based on the said attributes, the results imply that using these approaches is feasible. However, there are inherent difficulties in integrating worker attributes into a task assignment system. Certain attributes like demographics are self-reported by workers. Comprehensive personality tests are time-consuming and there is a possibility for workers to manipulate the outcome. Also, less competent crowd workers tend to overestimate their performance in self-assessments [25].

Previous work has also shown that it is possible predict task performance based on worker behaviour for worker pre-selection [24, 34, 60]. In content creation and information finding tasks, Gadiraju et al. [24] classify workers into five categories using behavioural traces from completed HITs. The study demonstrates that significant accuracy improvements could be achieved by selecting workers to tasks based on given categories. Rzeszotarski and Kittur [60] examined the way workers complete HITs by extracting user activity like mouse movements, scrolling activities, and key-strokes. Their model can successfully predict output quality in content generation, classification and comprehension tasks. Han et al. [34] reported a similar relationship between worker behaviour and content quality in annotating tasks. However, analysing worker behaviour observed over a considerable time period does not provide the utility we aim to achieve through brief cognitive tests.

2.3 Cognitive Ability and Tests

The compatibility between job requirements and the respective worker, and the agreement between job expectations of the worker and the job specifics are two key aspects of the Person-Job fit theorem [46]. This person-job match is known to result in numerous benefits in different work environments such as enhanced job performance, and satisfaction and motivation [19]. Therefore, organisations often seek to achieve a high person-job compatibility for their positions and use a wide variety of performance measures like cognitive ability, personality, general knowledge, emotional intelligence, and work experience [62].

Human cognitive ability has been long identified as an indicator of performance in education [8] and at work [5]. Psychological tests like Stroop [50], Simon [39] and Corsi Block [47] are often used to capture and measure the cognitive ability and are widely used in medical and psychological research [14]. Many of such tests have also been implemented online as test kits or collections like Test My Brain [28] and Cambridge Neuropsychological Test Automated Battery (CANTAB) [59]. In a study that uses Test My Brain, Germine et al. [28] show that it is viable to conduct cognitive tests on the web. Further, Crump et al. [10] conducted a study where crowd workers in MTurk platform were asked to complete cognitive tests such as Stroop, Flanker and Attention Blink. They show that results do not differ from lab-based studies and that it is feasible to use crowdsourcing platforms for such behavioural experiments.

In this study, we aim to use short online cognitive tests to capture the cognitive skills of crowd workers, and use the test outcome to predict their crowd task performance.

2.4 Impact of Cognitive Ability on Crowd Task Performance

Previous work by Eickhoff [20] and Alagarai Sampath et al. [1] indicate the possibility of using cognitive tests for crowdsourcing task assignment. The study by Eickhoff [20] investigates cognitive biases, a closely related trait to cognitive skills. Cognitive biases are known as systematic errors in thinking and can impact peoples everyday judgements and decisions. Literature also reports that cognitive skills can help people avoid cognitive biases [65]. The study shows that cognitive biases negatively impact crowd task performance in relevance labelling. Furthermore, Alagarai Sampath et al. [1] examined the cognitive elements in crowd task design. The study shows that reducing the demand for cognitive work, such as tasks involving visual search and working memory, could lead to higher overall task accuracy.

Goncalves et al. [29] first examined the possibility of using cognitive tests to predict the crowdsourcing task performance using a lab study. While the study reports promising results, it uses a set of time-consuming and paper-based cognitive tests from ETS cognitive kit [21] that are not practical for an online setting. A recent study by Hettiachchi et al. [36] investigates the effect of cognitive abilities on crowdsourcing task performance in an online setting. The work leverages the three executive functions of the brain (inhibition control, cognitive flexibility and working memory) [14] to describe and model the relationship between cognitive tests and crowdsourcing tasks. The study conducted on MTurk with the participation of 102 workers shows that there is a significant correlation between the cognitive test and crowdsourcing task performance. Further, they use multiple models to predict the task performance and show that a worker selection based on predicted scores could lead to better task accuracy.

Our work builds on this prior work as we aim to present an online dynamic task assignment framework that uses cognitive test results to estimate the worker performance, and assign the workers to suitable tasks.

3 STUDY

Next, we detail our experimental design starting with a description of the cognitive tests and crowdsourcing tasks used in the study. Then, we describe the proposed task assignment method, followed by details of the system architecture and study deployment.

3.1 Cognitive Tests

We use five cognitive tests similar to those used in a previous study by Hettiachchi et al. [36]. A description of each cognitive test is provided below, followed by Figure 1 which shows an example of each test. Results from these cognitive tests are used to inform worker task assignment.

3.1.1 ***Stroop*** *Test [50].* Stroop test is one of the classical cognitive tests that evaluate the human ability to overpower the prepotent response to words. In this test, participants encounter three types of trials (incongruent, congruent and unrelated). In incongruent trials, participants see the name of a colour displayed in another colour (*e.g.,* the word "blue" written in a "green" colour as shown in Figure 1). For congruent trials, the name of the colour matches the display colour. In unrelated trials, words displayed are non-colour words. In each trial, the participant needs to ignore the meaning of the word and respond to the colour of the word by pressing a key. Stroop effect states that people are less accurate and slower in incongruent trials when compared with congruent trials.

3.1.2 Eriksen ***Flanker*** *Test [22].* Similar to Stroop test, Flanker test also measures inhibition control but uses a different element. Here, we present 16 trials with two types of images that show five arrow symbols. Congruent trials show all arrows in same direction (*e.g.,* >>>>>) whereas incongruent trials show arrow in the middle in opposite direction (*e.g.,* <<><<). We ask participants to focus on and respond to the symbol in the centre. For the Flanker test, literature reports an effect similar to the Stroop test.

3.1.3 ***Task Switching*** *Test [53].* As shown in Figure 1, in the task switching test, participants see a letter and a number in one of the four squares in a 2×2 layout. In each trial, participants need to respond to one of the two questions; 'is the letter a vowel?' or 'is the number even?' depending on the position the stimuli appearing on the grid. Two types of trials are present in this test. Repeating trials let the participant answer the same question as the previous trial whereas switching trials force participants to change the question from the previous trial. There are 16 trials with 8 of each type.

3.1.4 ***N-Back*** *Test [55].* N-Back test measures the working memory of individuals by asking them to keep track of a series of stimuli. We use the 3-Back version of this test with 16 trials and letters appearing at each trail as shown in Figure 1. Participants are asked to decide if the current letter matches with the one they saw 3 trials ago.

3.1.5 Self-ordered ***Pointing*** *Test [58].* Pointing task tests participants ability to remember a sequence of recent actions. Here, we present 5 trials. In each trial, participants see 3 to 12 squares randomly distributed but identical in size. At any given time, a single square contains a reward. Participants are required to click one square at a time, without repeating until the reward is found. At each click, visual feedback indicates if the reward is found. The reward switches to a different square each time its found and the trial ends when the reward has shifted to all the squares in the trial.

Each cognitive test measures one of the three core executive functions of the brain as detailed in Table 1. *Inhibition control* is the conscious or unconscious restriction of a process or behaviour, particularly of impulses or desires. *Working memory* is the ability to hold information in memory and mentally work with it. *Cognitive flexibility* or Switching is the ability to adapt behaviours in response to changes in the environment and is often associated with creativity [14, 54].

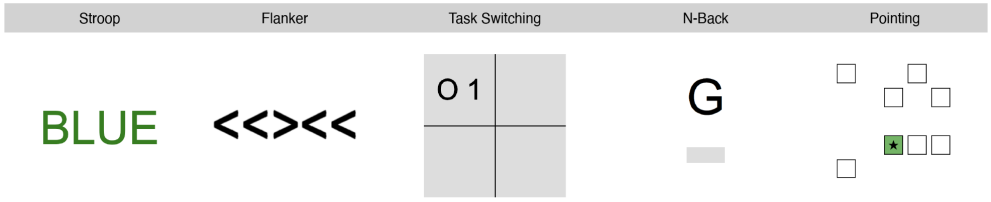

Fig. 1. Examples of each Cognitive Test

Table 1. Cognitive tests and primary executive function they measure [14]

Executive Function	Cognitive Test(s)
Inhibition Control	Stroop and Flanker
Cognitive Flexibility	Task Switching
Working Memory	N-back and Pointing

We provide instructions as well as an example before each test to aid workers. Apart from the pointing test, each trial in all tests is set to expire in 3.5 seconds. This important measure ensures that workers do not pause the test and prevents them from getting distracted while completing the test. We record accuracy and response time for each trial in the Stroop, Flanker, Task Switching, and N-Back tests. For the Pointing test, we gather the number of errors and the mean response time for trials in each round. Additionally, we record and use the trial type to calculate the test effect for Stroop, Flanker and Task Switching tests (*e.g.,* in Stroop test, the difference in accuracy between congruent and incongruent trials is called the Stroop effect related to accuracy).

3.2 Crowdsourcing Tasks

We chose four different crowdsourcing tasks for our experiment. These tasks have been carefully curated based on a crowd task taxonomy [26] and task availability [16] from prior work to be representative of typical tasks available in crowdsourcing platforms. Counting and sentiment analysis tasks were originally utilised by Goncalves et al. [30] and Goncalves et al. [29]. Each crowdsourcing task has multiple unique questions with varying complexity. Both sentiment analysis and counting tasks have 12 questions each while classification and transcription tasks have 9 questions. We also note that these tasks represent different answer types like multiple choice and text input. The tasks can be seen in Figure 2.

*3.2.1 Item **Classification**.* This is a multiple choice question with one or more possible correct answers. Each question contains a painting sourced from The Metropolitan Museum of Art[1] or Flickr[2] where all images are licensed for public use. Workers are given a list of four items and are asked to verify if the items are visible in the painting. Paintings depict a variety of styles that span into different continents. We use the following equation to calculate the accuracy for each question q with a set of A answers provided by a worker and a set of C correct answers. $Accuracy(q,A,C) = max\left[0, \sum_{a \in A} \frac{1}{|C|} \times \begin{cases} 1, & \text{if } a \in C \\ -1, & \text{otherwise} \end{cases}\right]$

*3.2.2 **Counting**.* The counting task presents workers the challenge of counting malaria-infected blood cells in a petri dish which also contain regular blood cells. Images we use in the task were generated using an algorithm to contain varying numbers of infected and regular blood cells. When

[1]https://www.metmuseum.org/art/collection
[2]https://www.flickr.com

workers provide a response a accuracy for each question q of this task with single correct answer c is calculated from $Accuracy(q,a,c) = max\left[0, 1 - \frac{|a-c|}{c}\right]$.

3.2.3 ***Sentiment Analysis.*** In this labelling task, workers determine the sentiment of a given sentence witch could be either 'positive', 'negative' or 'neutral'. Task contains two types of sentences. The sentiment of straightforward sentences like "My friends think the price is too expensive" can be easily classified. Other sentences like "Absolutely adore it when my bus is late." are more challenging due to context or language specifics like sarcasm. When a worker provides an answer a to a question q with a correct answer c, we use $Accuracy(q,a,c) = \{^{1,\ \text{if } a=c}_{0,\ \text{otherwise}}$ to calculate the accuracy.

3.2.4 ***Transcription.*** The transcription task presents workers with an image that contain several text elements. Workers require to recognise and type the text content in a provided text box. We use image segments from The George Washington Papers at the Library of Congress [64] [3]. Due to the time and individual variations in handwriting, selected images have varying complexity. To obtain the accuracy for each question q with a correct answer c and response a, we calculated Levenshtein distance (LD) [11] between the response string and the ground truth and used the equation $Accuracy(q,a,c) = max\left[0, 1 - \frac{2 \times LD(a,c)}{string\_length(c)}\right]$.

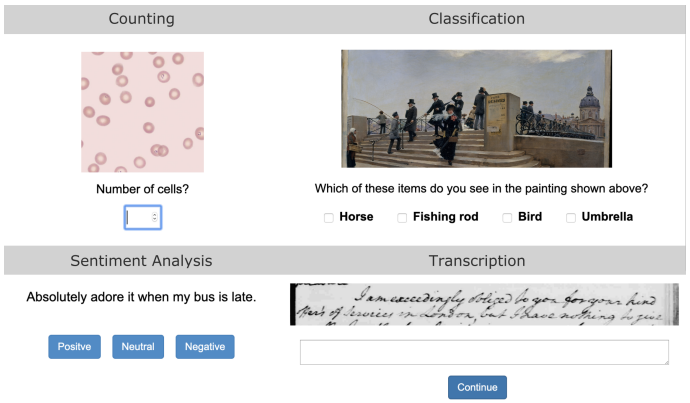

Fig. 2. Examples of Each Crowdsourcing Task

3.3 Task Assignment

3.3.1 Problem. Here, we define the task assignment problem we attempt to solve in this work. Assume that we have a set of tasks $T = \{t_1,..,t_k\}$ and a set of workers $W = \{w_1,..,w_m\}$ where $|T| = k$ and $|W| = m$. Each task t may contain an arbitrary number of questions. In order to maximise the overall quality of the data we gather, for each worker, we aim to assign the task t' where the worker is more likely to produce results of better quality.

The problem we attempt to address in this work is slightly different from question assignment in crowdsourcing, which is also often referred to as 'task assignment' (*e.g.,* [44, 67]). Crowd tasks usually contain several sub-tasks or questions in each task. For example, consider the case shown in Figure 3. There are three tasks (*e.g.,* triangle, square and circle) with each task having four questions. When a worker requests a task, the aim of the task assignment is to select the most suitable task (*e.g.,* circle). Once a task is selected, the question assignment determines the specific question(s) that should be allocated.

[3] https://www.loc.gov/collections/george-washington-papers/about-this-collection

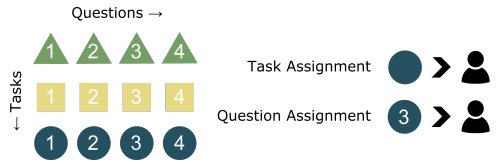

Fig. 3. Task assignment vs Question assignment

We propose two task assignment methods based on cognitive skills of crowd workers. In our first approach "CrowdCog-Assign" we aim to select and assign the optimum task for each worker as determined by our method. Our second approach "CrowdCog-Recommend" is a more relaxed approach where we provide workers with our task recommendation and let them select the task they want to work on. To help readers understand our proposed methods, an overview of the two proposed methods is provided in Figure 4. Here, green coloured elements in dashed line are exclusive to the CrowdCog-Recommend method while blue coloured elements in dotted line solely represent the CrowdCog-Assign method.

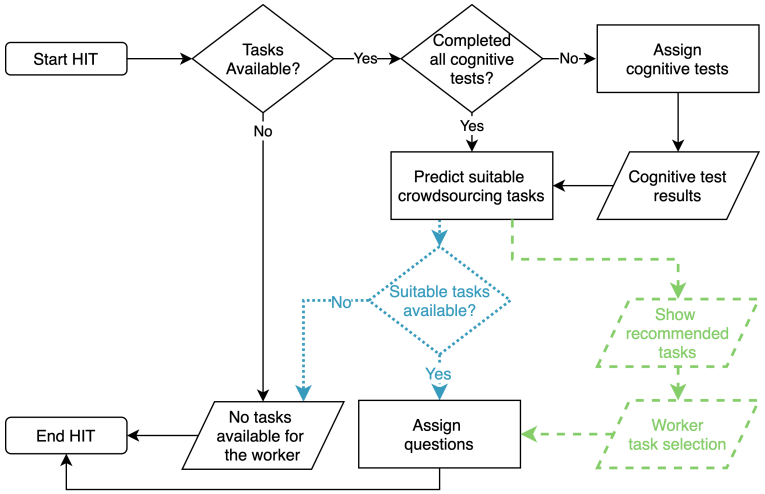

Fig. 4. Flow chart of CrowdCog-Assign in blue dotted line and CrowdCog-Recommend in green dashed line.

3.3.2 CrowdCog-Assign. We introduce a set of cognitive tests, $C = \{c_1,..,c_l\}$ where $|C| = l$ with each test measuring one of the three executive functions (inhibition control, cognitive flexibility, and working memory). We also define two parameters that determine the size of each task unit (*i.e.,* HIT in MTurk). The maximum number of cognitive tests to be included in each HIT, C_{HIT_MAX} and the maximum number of questions to be included in each HIT, Q_{HIT_MAX}.

For each task $t \in T$, we have a set of questions $Q_t = \{q_{t,1},..,q_{t,p}\}$. For each of these questions, we need to obtain an arbitrary number of votes or answers. For simplicity, here we assume all questions in all tasks require a Z number of votes. We also need to keep track of the number of votes or answers provided at a given moment. Lets define $V_t = \{v_{t,1},..,v_{t,p}\}$ where $v_{t,q}$ is the current number of votes or answers received for the question q in task t.

When a worker starts a HIT, we check if there are tasks available for the worker. This check is based on two steps. First, we obtain a list of questions that still need to be answered. For any task t, an available question $q_{t,j}$ is where $v_{t,j} < Z$. Second, for each question in the available question list, we remove questions the worker has already attempted based on the worker task completion history.

Then, we select the tasks that correspond to the remaining questions. At the end of this filtering step, we have a list of tasks T' that could be potentially assigned to the worker. If there are no tasks in the list, we end the HIT with a message to the worker.

Then we assign cognitive tests for the worker. Here, we keep track of the cognitive tests the worker has already completed C^w and first obtain a list of tests the worker has not completed yet ($C - C^w$). Then, we randomly assign a C_{HIT_MAX} number of tests or the total number of tests if $|C - C^w| < C_{HIT\_MAX}$. If the worker has already completed all the cognitive tests, we skip this step and directly move to task prediction. Following the cognitive test assignment, the worker will complete all the assigned tests and upon completion, the system will receive the results. Once we receive the cognitive test results we attempt to assign a task to the worker. Task prediction is based on the model and the relationship between cognitive tests and crowdsourcing tasks proposed by Hettiachchi et al. [36]. We use individual random forest models for each task with parameters number of trees set to 1000 and features selected at each split to 3.

Based on Table 1 and Table 2 and from prior work, we already know the set of cognitive tests (C^t) that a worker needs to complete in order for us predict the accuracy of that worker for a particular task t. For instance, if a worker has completed all the cognitive tests related to Cognitive Flexibility (*e.g.,* Task Switching), we can predict the accuracy for Transcription task using our model. Likewise, we predict the accuracy of all the available tasks for the current worker T^w. Then for each task t, if the predicted accuracy, $accuracy_{w,t}$ for worker w is higher than the pre-determined threshold $accuracy_t^0$, we consider that task as a candidate for the assignment. Finally, we select a task from the possible assignments. In our study, we select a random task from the candidate list to best replicate a real-life crowd-market scenario where workers would be allowed to attempt many tasks that they qualify for as based on the results of a common set of cognitive tests. Therefore using our model, we can find the task that should be assigned to the worker as detailed in the Algorithm 1.

Table 2. Relationship between Crowdsourcing Tasks and Cognitive Tests [36]

Crowd Task	Significant Features	Related Executive Functions
Classification	Pointing (Accuracy) Flanker (Resp. Time) Stroop (Accuracy)	Inhibition Control Working Memory
Counting	Flanker (Effect Accuracy) Pointing (Resp. Time) Stroop (Accuracy)	Inhibition Control Working Memory
Sentiment Analysis	Stroop (Resp. Time) Instructions (Resp. Time) Flanker (Effect Accuracy)	Inhibition Control
Transcription	Task Switching (Accuracy) Task Switching (Effect Accuracy)	Cognitive Flexibility

Following the task assignment, we select the questions to be assigned to the crowd workers. For this purpose, we also keep track of the number of answers still required for each question to avoid redundant labels. The worker then completes the assigned questions. As the final step, we collect the responses for questions, record them and mark the HIT as submitted in the Amazon Mechanical Turk platform.

$C^w \leftarrow$ Set of cognitive tests completed by worker w
$T^w \leftarrow$ Set of available tasks for worker w
$assignment \leftarrow$ The task assigned for the worker w
input : C^w,T^w
output : $assignment$

$possible\_assignments \leftarrow \emptyset$; $assignment \leftarrow \varnothing$;
foreach $t \in T^w$ **do**
  if $\forall c(c \in C^t \cap C^w)$ **then**
    $accuracy_{w,t} \leftarrow Predict(c)$;
    if $accuracy_{w,t} > accuracy_t^0$ **then**
      $possible\_assignments \leftarrow possible\_assignments \cup \{t\}$;
    end
  end
end
if $possible\_assignments$ *is not* $\emptyset$ **then**
  $assignment \leftarrow RandomSample(possible\_assignments)$
end

Algorithm 1: Task assignment based on cognitive test results

3.3.3 CrowdCog-Recommend. A key restriction in the proposed CrowdCog-Assign assignment strategy is that it does not let workers select the tasks they want to work on. While this may have a positive impact on the performance, in certain cases, a crowdsourcing platform might still prefer to provide workers with the flexibility of selecting their own tasks. To allow for this, we propose the CrowdCog-Recommend method.

For this approach, we follow a similar process as the CrowdCog-Assign method until tasks are predicted from cognitive test results. As we finish iterating over tasks in T^w, we return *possible_assignments* without selecting a single task (See Algorithm 1). Instead of assigning the task, here we present workers with our task recommendation and ask them to select the task they want to work on. After task selection, the rest of the process is identical to the CrowdCog-Assign method.

3.4 Study Conditions

The study was conducted under five conditions as described below.

- *Baseline*: In the baseline, workers select the task they want to work on and the questions are randomly assigned by the system. The baseline is comparable to the task assignment in a generic crowdsourcing platform like MTurk.
- *CrowdCog-Assign*: The proposed method where tasks are directly assigned based on the cognitive test performance and questions are assigned randomly.
- *CrowdCog-Recommend*: The proposed method with tasks recommended using cognitive test results. Workers see the recommendation but still have the liberty to choose any task. Questions are assigned randomly.
- *QASCA*: We compare with QASCA proposed by Zheng et al. [67]. Under QASCA, workers select the task but questions are assigned based on Expectation Maximisation. We chose QASCA as it has been shown to perform better when compared to four other methods CDAS [48], AskIt! [7], MaxMargin and ExpLoss.
- *History-based*: Under this method which uses historical data, task and question selection is similar to the baseline. However, workers are allowed to attempt tasks only if they have

previously completed 1000 HITs in the platform with an approval rate of 95% or above. This worker selection criteria is widely utilised by researchers and the literature reports a significant increase in data quality when selecting workers with a high approval rate and a high number of HITs completed [57].

We deployed all four tasks under these conditions. As QASCA is originally proposed for multi-label questions with a single correct answer, we were not able test transcription task which gathers text input. Also under QASCA, we had to transform the answers for counting tasks into three labels using a bracketing method as suggested in the prior work [56]. For classification task which contains multiple correct labels, we only considered a single option when evaluating with QASCA.

For CrowdCog-Assign and CrowdCog-Recommend conditions, each HIT contained a maximum of 2 cognitive tests ($C_{HIT\_MAX} = 2$) as we need results from at least two cognitive tests to make a task assignment or a recommendation (See Table 2). Each HIT also included a maximum of 3 questions ($Q_{HIT\_MAX} = 3$) to be consistent with the study design of prior work [67] and to ensure we can equally distribute all questions within a task (our tasks contain either 12 or 9 questions). For the evaluation, we set the threshold for task assignment ($accuracy_t^0$) at a modest 50% accuracy of each task. This threshold can be adjusted by the requester depending on the urgency of the data collection and available funds.

Each condition was deployed in MTurk at independent iterations. Each iteration was deployed during the same time window on weekdays. Using a qualification, we prevented any worker from attempting tasks in more than one condition. We only allowed workers from the United States and workers were compensated at the rate of $0.4 (USD) for each HIT. The payment was decided based on the time estimations gathered from our pilot study and the highest state minimum wage of the United States $13.25. Workers were compensated with a bonus payment of $ 0.2 (USD) for each cognitive test they completed in addition to the tasks. We ensured the bonus payment is issued for cognitive tests even when no tasks were assigned to the workers. For all conditions except history-based, we did not employ any additional worker selection criteria like approval rate. The research is approved by the ethics committee of our university. When participants accepted their first HIT from our study, they were also required to accept an informed consent form in order to continue the study.

We built our system primarily using Python (Django Framework). The system was hosted in a standalone server and workers accessed tasks through the external task function in MTurk. The experiment was presented to the worker through a popup window that automatically submits the HIT at the end. Several elements that allow for this seamless integration with the MTurk platform were extended from *PsiTurk*, an open platform for building experiments on MTurk [33]. For the creation of cognitive tests, we also used *jsPsych*, a JavaScript library for running behavioural experiments in a web browser [13].

4 RESULTS

In our study, a total of 574 workers completed 983 task units (HITs) across five conditions. Completed HITs accounted for 838 cognitive tests and 1,703 answers for crowdsourcing tasks. On average workers spent 2.95 minutes on HITs that contained crowd tasks and 2.98 minutes on HITs that contained both cognitive tests and crowd tasks. For the analysis, we use task accuracy as the primary evaluation metric which is calculated as described under the crowdsourcing tasks section (Section 3.2).

4.1 Cognitive Test Validation

Participant responses collected for the three cognitive tests can be validated using the difference in trial accuracy and response time between different types of trials. For example, a one-sample Wilcoxon signed rank test shows that the difference in accuracy ($M = 0.13$, $SD = 0.22$) between congruent and incongruent trials in Stroop test is significantly higher than 0 ($V = 3377.5$, $p < 0.001$) whereas

a one-sample t-test shows the difference in response time ($M = -196.68$, $SD = 258.67$) is significantly lower than 0 ($t(183) = -10.31$, $p < 0.001$). Similarly, the difference in accuracy ($M = 0.25$, $SD = 0.39$) was significantly higher than 0 ($V = 4761.5$, $p < 0.001$) and response time ($M = -97.20$, $SD = 236.93$) was significantly lower than 0 ($t(171) = -5.38$,$p < 0.001$) for the Flanker test. In the Task Switching test, difference in accuracy ($M = 0.01$, $SD = 0.20$) and response time ($M = -17.61$, $SD = 378.65$) between switching and repeating trials are not significantly different from 0 as opposed to the Stroop and Flanker tests. The difference in direction follows the findings from prior literature [22, 50, 53].

4.2 Task Recommendation

For tasks completed under the CrowdCog-Recommend condition, we analyse the difference in accuracy between two cases. First, under *No Recommendation*, workers attempt a task when there is no task recommendation given from the system. Second, under *Attempt Recommended*, workers attempt a task that was recommended by the system. Figure 5 shows that workers performed better when attempting recommended tasks when compared to other tasks. A Wilcoxon rank sum test shows that task accuracy for Attempt Recommended case is significantly higher when compared to the No Recommendation case ($W = 21034$, $p < 0.01$). We also note that workers were more likely to accept a recommendation. In our CrowdCog-Recommend setting, workers were presented with a task recommendation in 89 HITs. Workers opted to work on a recommended task in 61 HITs (68.53%).

Fig. 5. Accuracy and Standard Error for each task for the task recommendation conditions

4.3 Task Assignment

Under the CrowdCog-Assign setting, 239 unique workers initiated our HIT and 63 (35.80%) of them were assigned to one or more tasks. Out of 176 workers who were not assigned to tasks, 156 (88.60%) workers did not attempt more than a single HIT which includes only 2 cognitive tests. In Figure 6 we observe that as workers completed more cognitive tests, they were more likely to be assigned to a task. We validate this observation through a Chi-squared test ($\chi^2 = 85.39$,$p < 0.001$). Further, when considering workers who completed all five tests, 72% of them were assigned to at least one task.

Fig. 6. Variation in task assignment against the number of cognitive tests completed

4.4 Comparing CrowdCog to Other Methods

We analyse and compare the performance of proposed CrowdCog methods with three other conditions: baseline, QASCA and history-based method. For CogCrowd-Assign we also included the answers obtained under attempt-recommended of CogCrowd-Recommend.

We report a significant improvement in the accuracy of the workers compared to the baseline. As the study comprises of tasks accounting for both discrete and continuous accuracy values, our data does not pass the Levene's test for homogeneity of variance and Shapiro-Wilk normality test. Hence, we use Kruskal-Wallis rank sum test and report a significant difference in accuracy ($\chi^2 = 32.37$, $p < 0.01$, df = 4) among five conditions. Further, we conduct a post-hoc analysis via Dunn Test with p-values adjusted with the Benjamini-Hochberg method. Results show that when compared to the baseline, the accuracy is significantly higher in the CrowdCog-Assign method ($Z = -4.17$, $p < 0.01$) as well as in the CrowdCog-Recommend method ($Z = -2.51$, $p = 0.02$). While accuracy of CrowdCog-Assign method is significantly higher when compared to QASCA ($Z = -2.64$, $p = 0.02$), there is no significant difference in accuracy between history-based method and CogCrowd-Assign ($Z = 1.11$, $p = 0.27$). Figure 7 visualises the mean accuracy and standard error values for all the tasks across the baseline and proposed methods. Accuracy values are also summarised in Table 3.

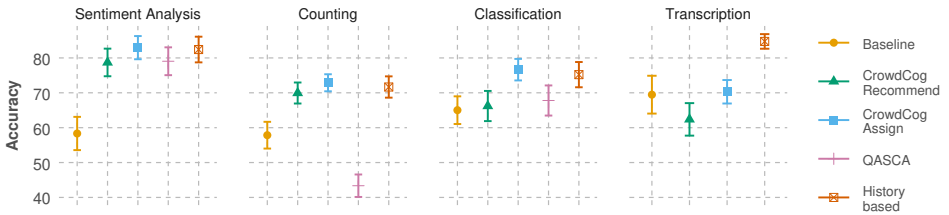

Fig. 7. Accuracy and Standard Error of tasks

Table 3. Task Accuracy across conditions

Condition	Baseline	CrowdCog Rec.	CrowdCog Assign	QASCA	History based
Sentiment Analysis	58.3	78.7	**82.9**	79.0	82.4
Counting	57.8	69.9	**72.9**	43.4	71.7
Classification	65.0	66.2	**76.6**	67.8	75.2
Classification [a]	64.2	78.0	**85.6**	71.6	80.0
Transcription	69.5	62.4	70.3	-	**84.7**

[a] Accuracy calculated considering only a single option to be comparable with QASCA

Figure 8 shows the mean response time in seconds for each task across three conditions. Although workers appear to be generally faster in our CrowdCog-Assign condition for most tasks, we do not observe any statistically significant difference in terms of response time across conditions.

To examine whether we have collected a sufficient number of responses for the tasks, we observe the variation in accuracy as we gather participant answers. Figure 9 shows that the accuracy is relatively stable after we accumulate 50% of the answers for sentiment analysis, counting, and classification tasks.

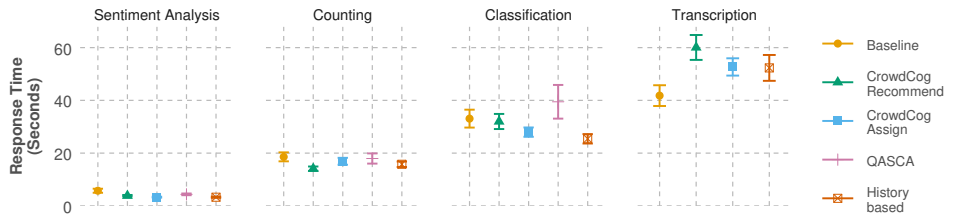

Fig. 8. Response Time and Standard Error of tasks

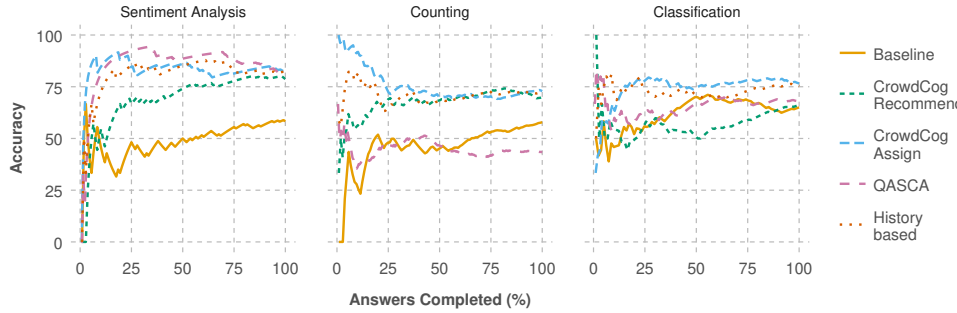

Fig. 9. Task accuracy against the answers completed

4.5 Cost Analysis

Our study included 42 questions across four tasks (Counting - 12, Classification - 9, Sentiment Analysis - 12, Transcription - 9) and we collected 9 answers for each question under different conditions. Here, in order to analyse the costs, we consider the order in which we received these answers and calculate the task accuracy by aggregating a varying number of answers. Figure 10 shows that fewer answers with CrowdCog-Assign method is sufficient to outperform the baseline with a larger number of answers. Next, we present a cost analysis where we only consider the first 3 answers for CrowdCog-Assign method and compare it against the baseline with 9 answers.

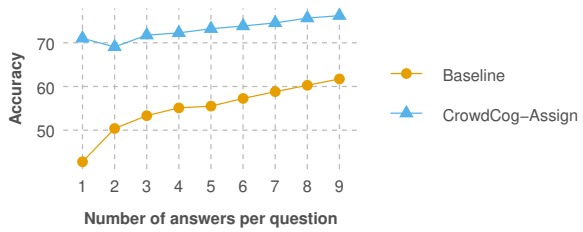

Fig. 10. Variation in task accuracy against the total number of answers aggregated

We show in Figure 11 that for all the tasks, the accuracy obtained from 3 answers per question under CrowdCog-Assign method is still higher than the accuracy from baseline with 9 answers per question. We calculate the total cost for 42 questions under the two conditions. First, under baseline, the cost is straightforward. As each answer costs \$0.13 (workers were payed \$0.4 for a HIT containing 3 questions), the total cost for obtaining 9 answers each for all the questions is $\$0.13 \times 9 \times 42 = \$49.14$.

Second, under CrowdCog-Assign method, the cost for all the answers would be $\$0.13 \times 3 \times 42 = \$16.38$. The additional cost for cognitive tests depend on the number of workers required for the task. We estimate the number of workers needed to obtain 3 answers, based on the number of

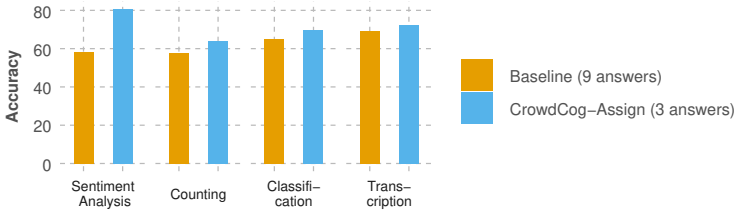

Fig. 11. Task accuracy with first 3 answers of each question from CrowdCog-Assign vs 9 from baseline

workers completed the study under this condition providing 9 answers for each question (174) and their cognitive test completion rates. The results show that 72.8% of workers completed only a single HIT (2 cognitive) tests, 6.3% completed two HITs (4 cognitive tests) and 20.9% completed three or more HITs (all 5 cognitive tests). Therefore, we determine the cost for cognitive tests $174 \times \frac{3}{9} \times (0.728 \times 2 + 0.063 \times 4 + 0.209 \times 5) \times \$0.2 = \$31.93$. Hence, the total cost for CrowdCog-Assign method adds up to \$48.31 in total. From Figure 11 and the calculated costs (Baseline \$49.14 and CrowdCog-Assign \$48.31), we show that the proposed CrowdCog-Assign method is capable of producing better results than the baseline at the same cost. While QASCA and history based methods do not result in additional costs, unlike CrowdCog, history based methods are not applicable for new workers and QASCA requires task specific calculations at each HIT submission.

5 DISCUSSION

5.1 CrowdCog Task Assignment

Crowdsourcing literature identifies task assignment in crowd platforms as one of the research foci [45]. Appropriate task assignment has many positive outcomes. From the perspective of a task requester, data quality can be increased while reducing the number of required labels, maximising cost-benefit. In the absence of task assignment, workers can find it challenging to locate appropriate tasks and tend to prioritise recently posted or new tasks, as well as tasks with the most number of HITs [9]. This also leads to requesters repeatedly posting the same task and flooding the platforms to attract workers [6]. If a platform is able to assign workers with compatible tasks, it will benefit workers by reducing the time and effort needed for task search and increasing worker satisfaction by achieving better person-job fit [19].

While numerous task assignment methods have been proposed, we note several shortcomings such as the inability to cater for a wide range of tasks (*e.g.,* [37, 44, 67]), and reliance on prior task records or external data (*e.g.,* [16, 24, 52, 60]). Concerning the validation of these previously introduced assignment methods, many evaluations are limited to synthetic data (*e.g.,* [4, 37]), one or two tasks (*e.g.,* [52]), or an offline analysis as opposed to online dynamic task assignment (*e.g.,* [24, 29, 36, 60]).

Our results indicate that when compared to the baseline (workers select the task without any recommendations), the proposed CrowdCog-Assign method (tasks assigned based on worker's cognitive test performance) produces significantly more accurate results. This increase in worker accuracy ranges from 5% to 20% across a variety of different task types. We also show that our method which works with new workers can achieve similar results compared to a widely used worker qualification that relies on historical data. In addition, the history-based method aims to restrict the available worker pool to a limited subset of workers who generally perform well across tasks. In contrast, we show that our method can successfully match workers to different types of tasks. Under CrowdCog-Assign, 72% of the workers who completed all five cognitive tests were assigned to at least one task.

We highlight that the proposed method is straightforward to implement and can be practised by both task requesters and platforms. However, a platform-level implementation could yield greater benefits. Once worker cognitive test results are captured, they could be utilised to assign many tasks. We highlight several factors that should be considered. First, as the cognitive ability of a worker could vary over time [17], cognitive tests should be repeated at a reasonable frequency. When repeating tests, the pool of tests would ideally consist of multiple tests for each executive function (*e.g.,* Stroop, go/no-go, Simon and many other tests for Inhibition Control [14]) as well as variants of the same test (*e.g.,* Stroop Test [50]) to ensure workers do not get familiarised with tests. Nevertheless, as cognitive tests include fast-paced time-restricted trials, workers would find it difficult to manipulate the outcomes [10] when compared to other task-independent approaches that could work without historical records such as demographics [63], personality tests [42] and self-assessments [25]. Second, when finding the relevant cognitive test for a particular task that does not relate to any of the tasks examined in our work, future researchers will have to identify the related executive functions of the task. They can replicate the approach detailed by Hettiachchi et al. [36] to build a hypothesis based on broad literature on human psychology. Alternatively (or in addition), a pilot implementation that includes three cognitive tests representing three executive functions can be used to determine what executive functions relate well to specific tasks. Once the relevant executive functions are identified, it is straightforward to determine the relevant cognitive test [14]. Third, an accuracy threshold needs to be set (see Algorithm 1) for each task before assigning tasks. This could be achieved via a pilot task set or using values based on our work. The threshold could also vary depending on the urgency of data collection. A lower threshold will result in an increased data collection rate but a lower accuracy increment as compared to the baseline.

Naturally, crowd task requesters are cautious of the additional costs that can be associated with more complex task assignment methods or quality control mechanisms [2, 35]. For the majority of common methods, such as gold standard questions and qualification tests, this additional cost is repeated for every new task. We supplement our study with a cost analysis to emphasise that cognitive tests could be incorporated in a crowdsourcing marketplace without increasing the potential costs. As shown in Figure 10, a reduction in the total number of answers required when applying our method compensates for the additional expenses required for cognitive tests. Further, when compared to the number of questions we have in our tasks (12 or 9), a typical crowd task has a sufficient number of questions [15, 41] to account for the additional amount requesters need to invest on cognitive tests.

5.2 Task vs Question Assignment

As the end goal of data quality improvement in crowdsourcing could be achieved through both task and question assignment, we argue that our comparison with question assignment methods is important. Question assignment methods also represent a large portion of rigorous frameworks proposed in the literature [12]. Based on the results of our study, we establish that the performance of our method is better or similar to the state-of-the-art question assignment methods. When considering the performance of the counting task, we observe that the task accuracy for QASCA is not significantly different from the baseline. Each question in the counting task has a single numeric input which we transformed into three groups using bracketing to apply expectation maximisation. This is the probable reason for the sub-par performance. Although prior work on QASCA suggests bracketing for handling questions with numeric input, they only experiment with multiple choice questions with a single correct label [67].

Another important consideration when using a real-time task assignment method is the impact on performance. If we deploy a sophisticated question assignment method such as QASCA, we need to carry out certain calculations at the end of each HIT which typically contains one or a few questions. This accumulates to a high demand for computational power when we consider the task completion rate in a standard crowdsourcing platform [15]. Therefore, unless the requester maintains

a third party resource that can calculate real-time scores, it can be quite challenging to implement a question assignment method like QASCA within a crowdsourcing platform. Our method provides a less computationally costly solution by reusing the worker cognitive test results for estimating performance for a variety of tasks.

Further, we note that our method could be used along with any question assignment method. For instance, a platform could implement our proposed method for task selection and use any of the question selection methods for question selection. While such a fine-grained task assignment implementation would be complex and computationally intensive, it could potentially increase the accuracy even further.

5.3 Task Recommendation

While task assignment aims to maximise the overall performance, it is important to consider potential negative consequences for the workers in terms of agency. In crowdsourcing, 'self-identification of contributors' [40] or workers' liberty to attempt a task they prefer is deemed important. Thus, task recommendation is often considered a more flexible alternative to task assignment [27]. Our work shows that the use of task recommendation based on cognitive skills still achieves significantly higher task performance when compared to the baseline. Prior attempts on task recommendation in crowdsourcing mainly rely on user-provided profile data, feedback collected from previous tasks [3], or worker task browsing history [66]. Also, Geiger and Schader [27] in a review of crowdsourcing task recommendation systems, identify the lack of an online analysis as a major drawback of the previous studies. In our study, we apply an online empirical analysis which shows that task recommendation based on workers' cognitive ability can lead to higher data quality when compared to a baseline of worker task selection.

In addition to the positive task recommendations applied in this paper, future work could potentially indicate negative task recommendations for tasks that are not recommended for a worker. This will allow workers to distinguish between tasks that are not recommended for them based on cognitive tests and tasks for which we are unable to make a prediction.

5.4 Limitations

We acknowledge several limitations of our study. First, many online task assignment frameworks often experiment with synthetic data to validate the proposed methods (*e.g.,* [4, 37, 61]). A handful of these studies have complemented the synthetic study results with a small scale real-time deployment on a platform like MTurk (*e.g.,* [44, 67]). However, because our results are based on cognitive tests, we only validate our method using a real-world deployment albeit with a high number of crowd workers. Unlike synthetic studies, real-world deployment limits our ability to extensively explore different parameter configurations. Second, we do not compare with any of the heterogeneous task assignment methods [4, 52]. This is mainly due to the incompatibility with our study setting and complexity in implementation of such proposed methods. However, we do compare with state-of-the-art question assignment methods.

6 CONCLUSION AND FUTURE WORK

In this paper, we study the heterogeneous task assignment problem through a novel and online assignment and recommendation method. We propose the use of short online cognitive tests for dynamic task assignment in a crowdsourcing platform across a variety of tasks. We built the CrowdCog system by integrating our novel task assignment and recommendation framework with MTurk. We evaluate the system using a real world study involving 574 crowd workers and 983 HITs across four tasks. Our study compares the proposed task assignment and task recommendation methods with a baseline generic task assignment and reports significantly higher task accuracy in both cases. We also show

that the proposed methods are comparable in improving worker's task accuracy when compared to state-of-the-art question assignment methods as well as a standard history-based qualification. At the same time, our method has a number of additional advantages, such as applicability to a variety of different tasks, not relying on historical performance data, and a better person-job fit which has been shown to lead to higher worker satisfaction [19].

Future work could explore a selection mechanism that takes into account the current task availability and cognitive test completion of the worker to further enhance the efficiency and productivity of the proposed method. Furthermore, once we have a list of eligible tasks for a worker, we randomly select a task from the list as opposed to the use of an optimised selection method. While this selection is less likely to impact the accuracy, an informed selection at this stage could further improve the efficiency of the data collection process. However, as both these enhancements dependent on various factors, future work in this domain will require a carefully crafted study design to account for the added complexity. In addition, a longitudinal study which investigates the frequency with which the cognitive tests should be repeated and the strategies for reusing cognitive tests will further strengthen the applicability of our findings.

REFERENCES

[1] Harini Alagarai Sampath, Rajeev Rajeshuni, and Bipin Indurkhya. 2014. Cognitively Inspired Task Design to Improve User Performance on Crowdsourcing Platforms. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. ACM, New York, NY, USA, 3665–3674. https://doi.org/10.1145/2556288.2557155

[2] Mohammad Allahbakhsh, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar. 2013. Quality Control in Crowdsourcing Systems: Issues and Directions. *IEEE Internet Computing* 17, 2 (2013), 76–81. https://doi.org/10.1109/MIC.2013.20

[3] Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2011. Towards Task Recommendation in Micro-task Markets. In *Human Computation Workshop at the Twenty-Fifth AAAI Conference on Artificial Intelligence*. AAAI Press, Palo Alto, California, USA.

[4] Sepehr Assadi, Justin Hsu, and Shahin Jabbari. 2015. Online assignment of heterogeneous tasks in crowdsourcing markets. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing (HCOMP'15)*. AAAI Press, Palo Alto, California, USA.

[5] Charles E. Bailey. 2007. Cognitive Accuracy and Intelligent Executive Function in the Brain and in Business. *Annals of the New York Academy of Sciences* 1118, 1 (2007), 122–141. https://doi.org/10.1196/annals.1412.011

[6] Michael S. Bernstein, David R. Karger, Robert C. Miller, and Joel Brandt. 2012. Analytic methods for optimizing realtime crowdsourcing. In *Proceedings of Collective Intelligence 2012*.

[7] R. Boim, O. Greenshpan, T. Milo, S. Novgorodov, N. Polyzotis, and W. Tan. 2012. Asking the Right Questions in Crowd Data Sourcing. In *2012 IEEE 28th International Conference on Data Engineering*. 1261–1264. https://doi.org/10.1109/ICDE.2012.122

[8] Erika Borella, Barbara Carretti, and Santiago Pelegrina. 2010. The Specific Role of Inhibition in Reading Comprehension in Good and Poor Comprehenders. *Journal of Learning Disabilities* 43, 6 (2010), 541–552. https://doi.org/10.1177/0022219410371676

[9] Lydia B. Chilton, John J. Horton, Robert C. Miller, and Shiri Azenkot. 2010. Task search in a human computation market. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*. ACM, 1–9.

[10] Matthew J. C. Crump, John V. McDonnell, and Todd M. Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE* 8, 3 (2013), 1–18. https://doi.org/10.1371/journal.pone.0057410

[11] Fred J. Damerau. 1964. A Technique for Computer Detection and Correction of Spelling Errors. *Commun. ACM* 7, 3 (1964), 171–176. https://doi.org/10.1145/363958.363994

[12] 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 Comput. Surv.* 51, 1, Article 7 (Jan. 2018), 40 pages. https://doi.org/10.1145/3148148

[13] 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. https://doi.org/10.3758/s13428-014-0458-y

[14] Adele Diamond. 2013. Executive Functions. *Annual Review of Psychology* 64, 1 (2013), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

[15] Djellel E. 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 (Florence, Italy) *(WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 238–247. https://doi.org/10.1145/2736277.2741685

[16] Djellel E. Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2013. 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* (Rio de Janeiro, Brazil) *(WWW '13)*. ACM, New York, NY, USA, 367–374. https://doi.org/10.1145/2488388.2488421

[17] Tilman Dingler, Albrecht Schmidt, and Tonja Machulla. 2017. 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, Article 47 (Sept. 2017), 15 pages. https://doi.org/10.1145/3132025

[18] Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are Your Participants Gaming the System?: Screening Mechanical Turk Workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. ACM, New York, NY, USA, 2399–2402. https://doi.org/10.1145/1753326.1753688

[19] Jeffrey R. Edwards. 1991. *Person-job fit: A conceptual integration, literature review, and methodological critique.* John Wiley & Sons.

[20] Carsten Eickhoff. 2018. Cognitive Biases in Crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM '18)*. ACM, New York, NY, USA, 162–170. https://doi.org/10.1145/3159652.3159654

[21] Ruth B. Ekstrom, Diran Dermen, and Harry Horace Harman. 1976. *Manual for kit of factor-referenced cognitive tests.* Vol. 102. Educational Testing Service, Princeton, NJ, USA.

[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. https://doi.org/10.3758/BF03203267

[23] Ju Fan, Guoliang Li, Beng Chin Ooi, Kian-lee Tan, and Jianhua Feng. 2015. iCrowd: An Adaptive Crowdsourcing Framework. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (Melbourne, Victoria, Australia) *(SIGMOD '15)*. ACM, New York, NY, USA, 1015–1030. https://doi.org/10.1145/2723372.2750550

[24] Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2018. Crowd Anatomy Beyond the Good and Bad: Behavioral Traces for Crowd Worker Modeling and Pre-selection. *Computer Supported Cooperative Work (CSCW)* (Jun 2018). https://doi.org/10.1007/s10606-018-9336-y

[25] Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. 2017. Using Worker Self-Assessments for Competence-Based Pre-Selection in Crowdsourcing Microtasks. *ACM Trans. Comput.-Hum. Interact.* 24, 4, Article 30 (Aug 2017), 26 pages. https://doi.org/10.1145/3119930

[26] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. 2014. A Taxonomy of Microtasks on the Web. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media* (Santiago, Chile) *(HT '14)*. ACM, New York, NY, USA, 218–223. https://doi.org/10.1145/2631775.2631819

[27] David Geiger and Martin Schader. 2014. Personalized task recommendation in crowdsourcing information systems – Current state of the art. *Decision Support Systems* 65 (2014), 3–16. https://doi.org/10.1016/j.dss.2014.05.007

[28] Laura Germine, Ken Nakayama, Bradley C. Duchaine, Christopher F. Chabris, Garga Chatterjee, and Jeremy B. Wilmer. 2012. Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review* 19, 5 (2012), 847–857. https://doi.org/10.3758/s13423-012-0296-9

[29] 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* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1023–1031. https://doi.org/10.1145/3041021.3055128

[30] 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* (Zurich, Switzerland) *(UbiComp '13)*. ACM, New York, NY, USA, 753–762. https://doi.org/10.1145/2493432.2493481

[31] Jorge Goncalves, Simo Hosio, Denzil Ferreira, and Vassilis Kostakos. 2014. Game of Words: Tagging Places through Crowdsourcing on Public Displays. In *Proceedings of the 2014 Conference on Designing Interactive Systems* (Vancouver, BC, Canada) *(DIS '14)*. Association for Computing Machinery, New York, NY, USA, 705–714. https://doi.org/10.1145/2598510.2598514

[32] Jorge Goncalves, Simo Hosio, Niels van Berkel, Furqan Ahmed, and Vassilis Kostakos. 2017. CrowdPickUp: Crowdsourcing Task Pickup in the Wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 51 (Sept. 2017), 22 pages. https://doi.org/10.1145/3130916

[33] 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 (01 Sep 2016), 829–842. https://doi.org/10.3758/s13428-015-0642-8

[34] Shuguang Han, Peng Dai, Praveen Paritosh, and David Huynh. 2016. Crowdsourcing Human Annotation on Web Page Structure: Infrastructure Design and Behavior-Based Quality Control. *ACM Trans. Intell. Syst. Technol.* 7, 4, Article 56 (April 2016), 25 pages. https://doi.org/10.1145/2870649

[35] Danula Hettiachchi, Zhanna Sarsenbayeva, Fraser Allison, Niels van Berkel, Tilman Dingler, Gabriele Marini, Vassilis Kostakos, and Jorge Goncalves. 2020. "Hi! I Am the Crowd Tasker" Crowdsourcing through Digital Voice Assistants. 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/10.1145/3313831.3376320

[36] Danula Hettiachchi, Niels van Berkel, Simo Hosio, Vassilis Kostakos, and Jorge Goncalves. 2019. Effect of Cognitive Abilities on Crowdsourcing Task Performance. In *Human-Computer Interaction – INTERACT 2019*. Springer International Publishing, Cham, 442–464. https://doi.org/10.1007/978-3-030-29381-9_28

[37] Chien-Ju Ho, Shahin Jabbari, and Jennifer Wortman Vaughan. 2013. Adaptive Task Assignment for Crowdsourced Classification. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 28)*. PMLR, Atlanta, Georgia, USA, 534–542. http://proceedings.mlr.press/v28/ho13.html

[38] Chien-Ju Ho and Jennifer Wortman Vaughan. 2012. Online Task Assignment in Crowdsourcing Markets. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (Toronto, Ontario, Canada) *(AAAI'12)*. AAAI Press, Palo Alto, California, USA, 45–51.

[39] Bernhard Hommel. 2011. The Simon effect as tool and heuristic. *Acta Psychologica* 136, 2 (2011), 189–202. https://doi.org/10.1016/j.actpsy.2010.04.011 Responding to the Source of Stimulation: J. Richard Simon and the Simon Effect.

[40] Jeff Howe. 2008. *Crowdsourcing : why the power of the crowd is driving the future of business* (1st ed.). Crown Business, New York, NY, USA.

[41] 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. *Proc. VLDB Endow.* 10, 7 (2017), 829–840. https://doi.org/10.14778/3067421.3067431

[42] 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* (Glasgow, Scotland, UK) *(CIKM '11)*. ACM, New York, NY, USA, 1941–1944. https://doi.org/10.1145/2063576.2063860

[43] 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* (Maui, Hawaii, USA) *(CIKM '12)*. ACM, New York, NY, USA, 2583–2586. https://doi.org/10.1145/2396761.2398697

[44] 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* (Chicago, Illinois, USA) *(SIGMOD '17)*. ACM, New York, NY, USA, 1447–1462. https://doi.org/10.1145/3035918.3064055

[45] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, Texas, USA) *(CSCW '13)*. ACM, New York, NY, USA, 1301–1318. https://doi.org/10.1145/2441776.2441923

[46] Amy L. Kristof. 1996. Person-organization fit: an integrative review of its conceptualizations, measurement, and implications. *Personnel Psychology* 49, 1 (1996), 1–49. https://doi.org/10.1111/j.1744-6570.1996.tb01790.x

[47] Muriel D. Lezak, Diane B. Howieson, David W. Loring, and Jill S. Fischer. 2004. *Neuropsychological assessment.* Oxford University Press, New York, NY, USA.

[48] Xuan Liu, Meiyu Lu, Beng Chin Ooi, Yanyan Shen, Sai Wu, and Meihui Zhang. 2012. CDAS: A Crowdsourcing Data Analytics System. *Proc. VLDB Endow.* 5, 10 (June 2012), 1040–1051. https://doi.org/10.14778/2336664.2336676

[49] Ioanna Lykourentzou, Angeliki Antoniou, Yannick Naudet, and Steven P. Dow. 2016. 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* (San Francisco, California, USA) *(CSCW '16)*. ACM, New York, NY, USA, 260–273. https://doi.org/10.1145/2818048.2819979

[50] Colin M. MacLeod. 1991. Half a Century of Research on the Stroop Effect: An Integrative Review. *Psychological Bulletin* 109, 2 (1991), 163. https://doi.org/10.1037/0033-2909.109.2.163

[51] 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* (Montreal, Québec, Canada) *(WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 843–853. https://doi.org/10.1145/2872427.2883070

[52] Kaixiang Mo, Erheng Zhong, and Qiang Yang. 2013. Cross-task Crowdsourcing. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Chicago, Illinois, USA) *(KDD '13)*. ACM, New York, NY, USA, 677–685. https://doi.org/10.1145/2487575.2487593

[53] Stephen Monsell. 2003. Task switching. *Trends in Cognitive Sciences* 7, 3 (2003), 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

[54] Jonas Oppenlaender, Elina Kuosmanen, Jorge Goncalves, and Simo Hosio. 2019. Search Support for Exploratory Writing. In *Human-Computer Interaction – INTERACT 2019*. Springer International Publishing, Cham, 314–336.

https://doi.org/10.1007/978-3-030-29387-1_18

[55] 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. https://doi.org/10.1002/hbm.20131

[56] Aditya G. Parameswaran, Hector Garcia-Molina, Hyunjung Park, Neoklis Polyzotis, Aditya Ramesh, and Jennifer Widom. 2012. CrowdScreen: Algorithms for Filtering Data with Humans. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (Scottsdale, Arizona, USA) *(SIGMOD '12)*. ACM, New York, NY, USA, 361–372. https://doi.org/10.1145/2213836.2213878

[57] 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 (01 Dec 2014), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y

[58] 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. https://doi.org/10.1073/pnas.90.3.873

[59] T. W. Robbins, M. James, A. M. Owen, B. J. Sahakian, L. McInnes, and P. Rabbitt. 1994. 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 (1994), 266–281. https://doi.org/10.1159/000106735

[60] Jeffrey M. Rzeszotarski and Aniket Kittur. 2011. 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* (Santa Barbara, California, USA) *(UIST '11)*. ACM, New York, NY, USA, 13–22. https://doi.org/10.1145/2047196.2047199

[61] Morteza Saberi, Omar K. Hussain, and Elizabeth Chang. 2017. An Online Statistical Quality Control Framework for Performance Management in Crowdsourcing. In *Proceedings of the International Conference on Web Intelligence* (Leipzig, Germany) *(WI '17)*. ACM, New York, NY, USA, 476–482. https://doi.org/10.1145/3106426.3106436

[62] Frank L. Schmidt and John Hunter. 2004. General mental ability in the world of work: occupational attainment and job performance. *Journal of personality and social psychology* 86, 1 (2004), 162. https://doi.org/10.1037/a0012842

[63] Aaron D. Shaw, John J. Horton, and Daniel L. Chen. 2011. Designing Incentives for Inexpert Human Raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (Hangzhou, China) *(CSCW '11)*. ACM, New York, NY, USA, 275–284. https://doi.org/10.1145/1958824.1958865

[64] George Washington. 1766. George Washington Papers, Series 5, Financial Papers: Copybook of Invoices and Letters, 1754-1766. https://www.loc.gov/item/mgw500003

[65] Richard F. West, Maggie E. Toplak, and Keith E. Stanovich. 2008. Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology* 100, 4 (2008), 930. https://doi.org/10.1037/a0012842

[66] Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. 2015. Taskrec: A task recommendation framework in crowdsourcing systems. *Neural Processing Letters* 41, 2 (2015), 223–238. https://doi.org/10.1007/s11063-014-9343-z

[67] 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* (Melbourne, Victoria, Australia) *(SIGMOD '15)*. ACM, New York, NY, USA, 1031–1046. https://doi.org/10.1145/2723372.2749430

[68] Mengdie Zhuang and Ujwal Gadiraju. 2019. In What Mood Are You Today? An Analysis of Crowd Workers' Mood, Performance and Engagement. In *Proceedings of the 10th ACM Conference on Web Science* (Boston, Massachusetts, USA) *(WebSci '19)*. Association for Computing Machinery, New York, NY, USA, 373–382. https://doi.org/10.1145/3292522.3326010

Received January 2020; revised June 2020; accepted July 2020