Task Routing and Assignment in Crowdsourcing based on Cognitive Abilities

Jorge Goncalves[1,2], Michael Feldman[3], Subingqian Hu[2], Vassilis Kostakos[1,2], Abraham Bernstein[3]
[1]School of Computing and Information Systems, The University of Melbourne, Australia
[2]Center for Ubiquitous Computing, University of Oulu, Finland
[3]Dynamic and Distributed Information Systems Group, University of Zurich, Switzerland
[1,2]firstname.lastname@oulu.fi, [3]lastname@ifi.uzh.ch

ABSTRACT
Appropriate task routing and assignment is an important, but often overlooked, element in crowdsourcing research and practice. In this paper, we explore and evaluate a mechanism that can enable matching crowdsourcing tasks to suitable crowd-workers based on their cognitive abilities. We measure participants' visual and fluency cognitive abilities with the well-established Kit of Factor-Referenced Cognitive Test, and measure crowdsourcing performance with our own set of developed tasks. Our results indicate that participants' cognitive abilities correlate well with their crowdsourcing performance. We also built two predictive models (beta and linear regression) for crowdsourcing task performance based on the performance on cognitive tests as explanatory variables. The model results suggest that it is feasible to predict crowdsourcing performance based on cognitive abilities. Finally, we discuss the benefits and challenges of leveraging workers' cognitive abilities to improve task routing and assignment in crowdsourcing environments.

Keywords
Crowdsourcing; cognitive abilities; worker performance; visual tasks; fluency tasks; task routing; task assignment; Kit of Factor-Referenced Cognitive Tests.

1. INTRODUCTION
In this paper we evaluate a mechanism that can enable matching crowdsourcing tasks to suitable crowd-workers based on their cognitive abilities. The prevailing view of individual workers as computational units ignores underlying mechanisms of cognition such as complex emotions, creativity, and high-order thinking [14]. This is of particular importance as crowdsourcing tasks are expected to gain complexity and become more prevalent over the years to come [38], as opposed to the current prevalence of simpler tasks in crowdsourcing markets. Already, platforms such as Upwork, Elance and CrowdSource promote tasks that require high level of expertise and diverse talents [53].

Here we argue that it is imperative for the crowdsourcing research agenda to focus on mechanisms to match complex tasks to suitable crowdworkers [14]. This is a non-trivial challenge due to the diversity in motivation, cognition, and error amongst workers [3]. Currently, quality control is typically implemented as post-hoc filtering of substandard answers and "smart" aggregation of the crowd contributions. A common technique is to adopt a Gold Standard [10], which entails the creation and inclusion of tasks that have known answers to the requested crowdsourcing job. Another approach is to analyse the extent to which workers agree with each other in their answers [4,17,28]. However, these post-hoc approaches assume that the cognitive diversity of workers assigned to a given task cannot be controlled.

Our work proposes an alternative *a priori* approach. We develop a mechanism that identifies the cognitive abilities of workers, which in turn can enable a more effective assignment of tasks to suitable workers. Here, we only focus on visual and fluency tasks, as they are commonplace in crowdsourcing marketplaces, and the relevant cognitive abilities are known to vary considerably across the population [42,44].

Our study shows that for these types of tasks, worker performance can be predicted both by the Kit of Factor-Referenced Cognitive Tests [13], as well as performance in a sample set of visual and fluency-based crowdsourcing tasks that we developed. Our work shows that it is possible to reliably measure crowd-workers' cognitive skills, which can then be used to assign them to more suitable tasks where it would be expect them to perform well.

WWW 2017 Companion, *April 3-7, 2017, Perth, Australia.*
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3055128

2. RELATED WORK
2.1 Effect of Cognitive Abilities on Performance
The use of cognitive abilities as substantial factor for determining the work performance of individuals has been explored and validated in numerous studies within several research communities (*e.g.*, [1,30,37,41,58]). In particular, these studies explore the relationship between cognitive characteristics and task performance, highlighting the importance of cognitive abilities in predicting individual differences in job performance. These studies consistently point out that high cognitive abilities of workers lead to better job performance, and vice-versa [11,37]. Human behaviour may be represented as a mixture of personal factors, behaviour presets, and the social impact modifications due to the surrounding environment. Hence, these factors may have critical effect on human functioning, and therefore play a key role in human based systems [7].

Further, there are several widely used cognitive theories such as the Social Cognitive Theory [2], the Personal Construct Theory [52], the Cognitive Load Theory [43], or the Cognitive Dissonance Theory [46], that highlight the importance of considering cognitive elements within HCI research. The increased understanding that human behaviour needs to be reviewed as a conglomerate of cognitive and social processes has driven the increased adoption of these theories [14]. Several

studies within this area have focused on expertise assessment research, exploring cognitive and other aspects for intelligent expert-job matchmaking. For instance, Lee *et al.* [34] developed cognitive models aimed at measuring worker expertise based on the differences between responses. However, this approach works best when there is a large number of workers as it relies on cross-examination of as many responses to the tasks as possible to improve reliability.

The theoretical foundations of our study are based on the well-established Person-Environment fit theory [5], but in particular the Person-Job fit domain [32]. Person-job fit is defined as the compatibility between a person's characteristics and those of a specific job [32]. The effect of cognitive abilities on job performance has been broadly discussed in series of studies [23,56]. Specifically, previous work has investigated the impact of Person-Job fit on performance and strain [6]. Their findings showed that a high level of person-job cognitive style misfit substantially affects performance and strain. However, we note that there are a few studies that did not establish a significant link between cognitive abilities and performance [33,48]. Our work builds upon the preliminary exploration by Feldman and Bernstein [14] aimed at investigating this connection in the crowdsourcing domain to potentially exploit it for task assignment and routing purposes. Here, we extend their work by investigating different types of crowdsourcing tasks, controlling more carefully the background of participants and by strictly following the instructions given by the ETS factor-referenced cognitive tests kit in terms of time given to complete each individual task.

2.2 Task Routing and Assignment

The current *modus operandi* in the majority of crowdsourcing platforms is that the central authority of the system coordinates the assignment process. However, a number of techniques have been proposed in literature with the aim of improving task-routing and assignment in crowdsourcing. For instance, previous work has proposed using dynamic participant recruitment with characteristics that vary in space and time, while aiming to minimize the sensing cost without sacrificing coverage [35]. In another example, Shirani *et al.* [49] use a two-step process in the assignment of users for participatory text documentation. The first step, termed viewpoint selection, entails the selection of a minimum number of points in the urban environment from which the texture of the entire urban environment can be collected/captured. At the second step, called viewpoint assignment, the selected viewpoints are assigned to participants depending on a number of constraints (*e.g.*, restricted available time, starting point, and destination). Similarly, Reddy *et al.* [45] study a recruitment framework to identify appropriate workers based on past information to estimate their geographic and temporal availability.

In an online crowdsourcing scenario, Ho *et al.* [22] explore worker assignment to multiple and heterogeneous tasks based on a two-phase exploration-exploitation algorithm aimed at describing workers based on certain skill levels. In follow-up work, the authors explore allowing workers to specify upfront the maximum number of tasks they are willing to complete [21]. The skill levels of the workers are considered known (offline approach) or they become known gradually by following a learning process (online approach). Based on these skills, the platform then performs workers-to-task assignment. Another approach entails the use of Matrix Factorisation to predict a crowd worker's accuracy on new tasks based on previous performance by means of collaborative filtering [29]. Other proposed principles and methods for task routing aim to harness people's abilities to jointly contribute to a task and route the task onwards [60].

In the context of mobile social networks, Xiao *et al.* [59] explore the notion of cooperative assignment of tasks. In their study, they allow workers already assigned to tasks to reassign them to others in an attempt to minimize the task idle time. Mao *et al.* [36] present the construction of predictive models of engagement in volunteer crowdsourcing based on different sets of features that describe user behaviour. Finally, Horowitz and Kamvar [24] propose Aardvark, a platform that leverages social networks to route questions to suitable users. They allow a user to ask questions in natural language, which the system interprets and automatically routes to appropriate individuals in the user's social graph based on topic expertise, connectedness and availability.

These studies offer a number of different techniques to deal with task assignment and routing. However, many of these approaches do not take worker skills into account, while others do not offer a dominating strategy to deal with task assignment with consistent performance. Here, we explore the potential of using workers' inherent cognitive abilities as the primary factor for suitable task assignment and routing. For measuring cognitive abilities in our experiments we have chosen to use some of the Factor-Referenced Cognitive Tests constructed by the ETS (Educational Testing Service) [13].

3. STUDY

We conduct a lab study following a within-subjects design, as opposed to leveraging an online crowdsourcing platform. Our decision is based on a number of different factors: 1) it allowed us to follow precisely the instructions set by the well-established Kit of Factor-Referenced Cognitive Tests, which we use to measure participants' cognitive abilities, 2) to guarantee the understandability of each test by the worker, 3) to avoid other issues such as collusion or worker distraction, which can affect the results.

The Kit consists of 72 factor-referenced cognitive tests for 23 factors and aim to serve as a measurement for cognition dimensions. The kit of tests was originally published in 1976 and has gained validity and reliability across disciplines with the passing of time. The tests have been used in various domains such as multimedia learning, Alzheimer's disease research, decision-making, or human spatial cognition [1,37,51,58]. As our experiment relies on visual and fluency cognitive abilities, it is important to note that previous work has highlighted the appropriateness of these cognitive tests as a measure to cognitive ability for these two factors (*e.g.*,[9]).

3.1 Cognitive Abilities Tests

In our experiment we included 8 cognitive tests, 4 to measure cognitive abilities in visual tasks and another 4 to measure cognitive abilities in fluency tasks. The visual cognitive tests included the Hidden Patterns, Hidden Figures, Paper Folding, and the Surface Development tests. The fluency cognitive tests included Rewriting, Scramble Words, Making Sentences, and Arranging Words tests.

Due to time limitations, whilst ensuring internal consistency, we only took the first part of each test. Please refer to the ETS manual for details on each of these tests [13].

3.2 Crowdsourcing Tasks

The crowdsourcing tasks used in our study are similar to many of the typical crowdsourcing tasks found on crowd labour markets. In total, we developed 8 different crowdsourcing task categories. Table 1 provides an overview of the used tasks, including the number of unique tasks available and their description. The visual crowdsourcing tasks are adapted from previous work by Feldman & Bernstein [14], while the fluency tasks are designed specifically for this study. The Text Distortion task pertains to both visual and fluency factors. Further, it can be argued that Riddles pertain more to rational thinking than fluency cognitive abilities. However, we wanted to investigate to what extent fluency can play a role in solving crowdsourcing tasks that require rational thinking upon reading a textual description.

Table 1. Task categories, number of unique tasks available, and description of each task.

	Task Category	Unique Tasks Available	Description
Visual	Distance Evaluation	16	Evaluate which of two buildings is closer
	Height Evaluation	13	Evaluate which of two buildings is taller
	Item Recognition	16	Ascertain if certain items are shown in a picture
	Item Classification	25	Classify depicted bird into one of four types
	Text Distortion	12	Restore a distorted sentence akin to a captcha
Fluency	Proofreading	16	Find and correct mistakes within a sentence
	Sentiment Analysis	16	Classify the polarity of a given sentence
	Riddles	16	Solve a given riddle

All task categories include questions that vary in their complexity and are designed to cover different aspects of visual perception and fluency. This variance of complexity is important, as a person's cognitive abilities become more important in conducting time-consuming, complex tasks (rather than short term micro-tasks) [55]. The complexity for each task varied based on different parameters, such as:

- **Distance and height evaluation**: Different building perspectives taken from Google Maps that are simpler or more difficult to evaluate.
- **Item recognition:** Number of items in the picture (1 to 3) and degree of concealment of the items.
- **Item classification:** Distance of bird to the camera, and different bird perspectives and poses.
- **Text distortion:** Length of the sentence, used words and degree of distortion.
- **Proofreading:** Four sentences without any errors. The remaining sentences (12) included errors such as typos, incorrect annotation, properly spelled but incorrect word for the context, incorrect order of words and duplicate words.
- **Sentiment analysis:** Eight "straightforward" sentences (*e.g.*, "I hate it when she acts like that") and eight challenging sentences (that are also challenging for sentiment analysis tools), in which each one is based on one of four factors [8]:
    - *Context*: a sentence that contains a sentiment word that has opposite connotation depending on context (*e.g.* "The only downside of this restaurant is that it charges me too little for its service").
    - *Sentiment ambiguity*: a sentence that contains a positive or negative word, but does not express any sentiment (*e.g.* "Can you recommend a good tool I could use?") or sentences without sentiment words that express a particular sentiment (*e.g.* "This browser uses a lot of memory").
    - *Sarcasm*: a positive or negative sentiment word can switch sentiment if there is sarcasm in the sentence (*e.g.* "I'm so pleased road construction woke me up with a bang").
    - *Contronyms*: sentence that contains a word with two meanings that can change sentiment depending on the language used. This is often seen in slang, dialects, and language variations (*e.g.* "Their new album is so sick").
- **Riddles:** Several different riddles ranging from simple (*e.g.* "Which weighs more, a pound of feathers or a pound of bricks?") to complex (e.g., "I'm simple for a few people, but hard for them to hear, I live inside of secrets, I bring people's worst fears. What am I?").

3.3 Participants and Procedure

We recruited twenty-four participants from mailing lists of our university and social media (12 males, 12 females; ages: 18-36 years old, M = 26.3). Participants had a diverse range of educational backgrounds in order to increase the likelihood of participants having a diverse range of cognitive abilities. One third of the participants had a Natural Sciences background (*e.g.*, Biology, Ecology, Mathematics), another third had a Technical Sciences background (*e.g.*, Computer Engineering, Electrical Engineering, Wireless Communication) and the final third had a Social Sciences background (*e.g.*, Anthropology, Linguistics, Business). Each participant was paid 40 Euros for participating and the experiment (including intake, training, and data collection) lasted between 90 to 120 minutes per participant.

Participants arrived at our lab and they were initially briefed on the tasks they had to complete. We then recorded their personal information (age, gender, background) and proceeded with the experiment.

The experiment consisted of two stages:

- Measurement of visual and fluency cognitive abilities with the use of selected Factor-Referenced Cognitive Tests (Figure 1 left), and
- Completion of visual and fluency-based crowdsourcing tasks (Figure 1 right). To avoid learning effects the order of the cognitive tests and crowdsourcing tasks were counterbalanced.

Regarding the cognitive abilities measurements stage, participants were shown an instructions page before completing each individual test. Beyond explaining in detail what is required in a particular test, the instructions page also provides a few examples of acceptable answers and few training subtests in order for the participant to become accustomed with the process. Once the participants were comfortable with the instructions, they then

proceeded to complete that particular test. Following instructions given by the manual of the Factor-Referenced Cognitive Tests, a researcher kept strict timing using a handheld timer. Each test is designed with its own recommended time limit.

During the crowdsourcing stage, participants were instructed to complete the tasks on a desktop computer. The initial page asked participants to input their details (participant id, age, gender, background) to enable cross-referencing between the two datasets. As with the other stage, participants were shown an instructions page before each task, which included some examples. In addition, the Item Classification task (identify the correct bird in a picture), started by showing several pictures of each of the 4 bird types along with textual description of their characteristics. We recorded the participants' information and answers to each task, as well as the correct answer for easy verification.

Finally, we conducted a short semi-structured interview aimed at assessing participants' opinions on the different type of tasks. We asked participants to rank the crowdsourcing tasks from the least to the most challenging, while also indicating why they had more trouble with some tasks than others. We also enquired if, in general, they had a preference between visual or fluency tasks.

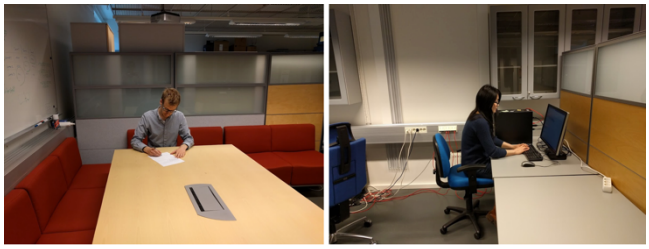

Figure 1. Left: Participant completing one of the ETS cognitive ability tests. Right: Participant completing one of the crowdsourcing tasks on a desktop computer.

4. ANALYSIS AND RESULTS

4.1 Prediction of Crowdsourcing Performance

For each participant, we calculated their average performance on the crowdsourcing tasks and the cognitive tests. A Pearson correlation test showed a significant positive correlation between the two variables ($r = 0.89$, $p < 0.01$), which can be visualised in Figure 2. This suggests that cognitive skills coincide with the performance on crowdsourcing tasks and can be leveraged for performance prediction.

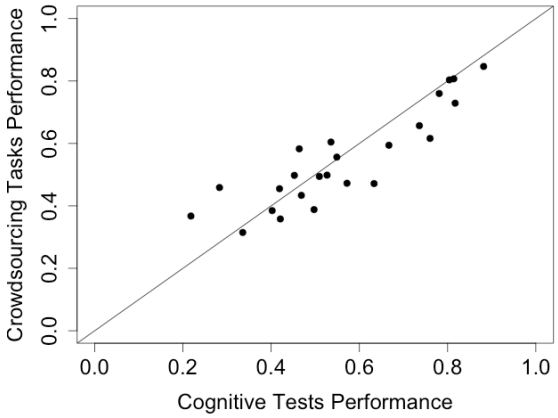

Figure 2. The average performance in crowdsourcing tasks plotted against average performance in the cognitive tests.

To investigate differences between individual participants, we also plotted their performance on visual and fluency cognitive tests (Figure 3 top), as well as performance on visual and fluency crowdsourcing tasks (Figure 3 bottom). To verify if educational background had an effect on performance, we colour-coded participants based on this factor. Both plots in Figure 3 show a wide distribution of participants at different levels of performance suggesting cognitive diversity, and there are no discernible clusters amongst those with similar backgrounds. The overall higher performance on the visual tests and tasks when compared to their fluency counterparts can be partly explained by the fact that while all participants were proficient in the English language, none of them were native English speakers. We discuss other further reasons that may explain this difference in the interview results. Overall accuracy of each task is 70% for building related tasks (76% distance and 64% height), 83% for Item Recognition, 42% for Item Classification and 74% for Sentiment Analysis.

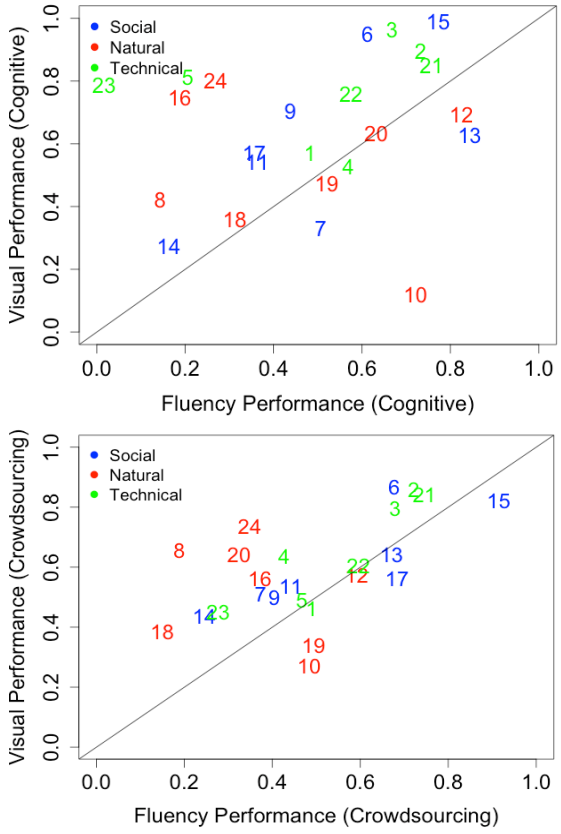

Figure 3. Top: Participant IDs and their performance in visual cognitive tests plotted against performance in fluency cognitive tests. Bottom: Participant IDs and their performance in visual crowdsourcing tasks plotted against performance in fluency crowdsourcing tasks.

Next, we built predictive models for the crowdsourcing tasks performance using the performance on the cognitive tests as explanatory variables. We constructed two predictive models: Beta and Linear regression (Table 2). Beta regression is typically used to model rates and proportions, and is suitable when the response variable is continuous and restricted to a range between 0 and 1. As our response data is beta distributed and mostly right skewed, we apply beta regression that has been shown to be appropriate in such cases [15]. We compare the results of the beta regression with the more widely used linear regression model as a

benchmark. Even though the assumption of normality does not hold for a linear regression, previous work has shown linear regression is robust to normality assumptions [54,57]. Table 2 provides the evaluation of two models in terms of Mean Average Error (MAE) and Root Mean Square Error (RMSE), and correlation between observed and predicted values. For each model, we calculated optimistic error based on the model trained on the whole dataset, and the error based on the leave-one-out (LOO) cross-validation. LOO cross-validation is a method to evaluate prediction error by iterative training of the model with all but one sample and then evaluating the prediction error on the sample that has not been used for model training. The LOO errors of beta and linear regression are very similar. Average MAE_{LOO} of Beta regression is 0.23, while Average MAE_{LOO} of linear regression is 0.225 with almost identical variance among the crowdsourcing tasks. The height evaluation and proofreading crowd tasks have the highest MAE and RMSE error rates, ranging between 0.259 and 0.27 in MAE, and between 0.321 and 0.345 in RMSE. The best predicted crowd tasks are item classification and sentiment analysis with almost identical MAE_{LOO} error rate of 0.197, correlation above 0.5 and the lowest RMSE errors among the tasks. Overall, the average correlation between the predicted and the observed performance ranged between 0.35 for the LOO and 0.76 for the optimistic errors. The similarity in the results provides additional evidence to the robustness of the presented prediction errors.

Table 2. Evaluation of the two prediction models (MAE and RMSE), and correlation between observed and predicted values.

		Error	Distance Evaluation	Height Evaluation	Item Classification	Item Recognition	Text Distortion	Sentiment Analysis	Proofreading	Riddle	Mean
Beta regression	**LOO**	MAE	0.235	0.27	0.229	0.197	0.221	0.216	0.246	0.233	0.231
		RMSE	0.266	0.345	0.296	0.255	0.268	0.275	0.301	0.289	0.287
		Correlation	0.352	0.072	0.332	0.53	0.491	0.474	0.195	0.388	0.354
	Optimistic	MAE	0.147	0.155	0.149	0.115	0.134	0.115	0.151	0.151	0.140
		RMSE	0.169	0.181	0.195	0.151	0.166	0.138	0.187	0.2	0.173
		Correlation	0.758	0.686	0.696	0.834	0.78	0.845	0.676	0.675	0.744
Linear regression	**LOO**	MAE	0.249	0.23	0.24	0.203	0.21	0.196	0.259	0.22	0.226
		RMSE	0.29	0.294	0.303	0.247	0.253	0.254	0.321	0.271	0.279
		Correlation	0.218	0.216	0.311	0.549	0.522	0.508	0.117	0.398	0.355
	Optimistic	MAE	0.144	0.137	0.143	0.12	0.128	0.109	0.146	0.143	0.134
		RMSE	0.163	0.179	0.183	0.147	0.161	0.134	0.175	0.184	0.166
		Correlation	0.774	0.691	0.73	0.844	0.801	0.855	0.717	0.721	0.767

The LOO method generates nearly unbiased estimates of the prediction error for small samples, however it can also lead to high variability of the results. Therefore, we conducted ordinary bootstrapping as this method gives an estimate of errors with low variability, but with possible bias of the errors [12]. Bootstrapping is a statistical technique to resample data from the existing sample in order to improve the statistical power of the results, and is both useful and commonly used when the predictions are based on a relatively small sample [12]. We applied bootstrap with 20 replications where the chance to draw each sample follows a binomial distribution and resulted in 480 observations (Table 3).

Table 3. Prediction table after applying bootstrapping.

		Error	Distance Evaluation	Height Evaluation	Item Classification	Item Recognition	Text Distortion	Sentiment Analysis	Proofreading	Riddle	Mean
Beta regression	**LOO**	MAE	0.149	0.156	0.144	0.12	0.144	0.114	0.159	0.167	0.144
		RMSE	0.171	0.183	0.194	0.154	0.174	0.138	0.19	0.211	0.177
		Correlation	0.771	0.656	0.687	0.831	0.759	0.852	0.663	0.609	0.729
	Optimistic	MAE	0.147	0.153	0.141	0.118	0.142	0.112	0.157	0.164	0.142
		RMSE	0.168	0.179	0.191	0.151	0.171	0.135	0.187	0.208	0.174
		Correlation	0.778	0.672	0.698	0.838	0.767	0.858	0.674	0.62	0.738
Linear regression	**LOO**	MAE	0.146	0.14	0.141	0.12	0.138	0.111	0.15	0.156	0.138
		RMSE	0.164	0.179	0.184	0.149	0.169	0.135	0.179	0.195	0.169
		Correlation	0.785	0.666	0.715	0.844	0.782	0.861	0.699	0.664	0.752
	Optimistic	MAE	0.144	0.138	0.139	0.118	0.136	0.109	0.147	0.153	0.136
		RMSE	0.162	0.176	0.18	0.146	0.167	0.132	0.176	0.192	0.166
		Correlation	0.793	0.68	0.727	0.85	0.789	0.867	0.712	0.676	0.762

4.2 Interview Results

We asked participants to rank the crowdsourcing tasks from the least (1) to the most (8) challenging. The task that was identified as the least challenging was Sentiment Analysis (average rank: 1.37), followed by Item Recognition (1.63), Distance Evaluation (2.74) and Item Classification (3.32). The participants identified the text distortion as the most challenging task (5.63), followed by Riddles (5.21), Proofreading (4.74) and Height Evaluation (3.37). One reason that may have contributed towards text Distortion being considered the most challenging task overall is the fact that it required a mixture of visual and fluency skills. This sentiment was expressed by several participants:

"Text distortion was very challenging for me. The text was very messy and difficult to the eye. It was hard for me to write sentences that made sense." - P1

"I had issues deciphering what was written in the Text Distortion task. Then when I got some words it was hard to create a logical sentence." - P6

Several participants (N=14) reported preferring and being more confident when completing the majority of visual tasks. These participants felt that the visual tasks were easier and played more towards their strengths:

"The visual based tasks were easier for me as I like extracting information from pictures." - P24

"I do not really write a lot in my daily life, so the visual tasks were just more natural and easier for me." - P20

"I preferred the visual based tasks as they are more interesting to me and I did not have to think as much when compared to the fluency tasks." - P18

This sentiment was reflected on the results of Figure 3 right, as these three participants scored better on the crowdsourcing visual tasks when compared to the fluency tasks. Contrarily, 8 participants preferred completing the fluency based tasks due to perceived ambiguity in the visual based tasks or lack of self-reported skills to perform well:

"I found the fluency based tasks easier as they were type-and-go. Visual tasks were hard to grasp for me, since you have to do everything in your mind as opposed to typing." - P19

"I preferred the fluency tasks, since the visual tasks were quite ambiguous in many cases and I ended up having to guess some of them." - P15

"I have always had problems with depth perception, so the building related tasks were difficult for me." - P10

Similar to those that preferred visual based tasks, these participants performed better when completing fluency-based tasks (Figure 3 right). The other 2 participants reported not having a preference. This difference in opinion amongst our participants confirms the cognitive diversity of our sample. However, the imbalance between the number of participants that either preferred the visual or the fluency based tasks can also further explain the differences in performance between these two types of tasks (higher overall performance on the visual tasks) as shown in Figure 3 right.

Finally, several participants stated the importance of having tasks that you feel capable of completing and how this can influence their performance:

"I would just get frustrated if I had to do tasks that I am not good at." - P2

"I could do these visual tasks for hours without a problem, but not the fluency ones." - P8

"If I was given tasks that I felt I was bad at, I would probably give up early or just not put a lot of effort in." - P9

5. DISCUSSION

As the complexity of crowdsourcing tasks increasingly grows, the need for labour markets and task requesters to improve their task assignment practices will become more important, particularly because this trend is expected to continue in the following years [38]. Also, cognitive abilities are likely to become more relevant in completing complex tasks as opposed to simpler tasks [55]. Our results show that by measuring workers' cognitive abilities we can predict their performance in typical crowdsourcing tasks, as suggested in a preliminary study by Feldman and Bernstein in an online scenario [14].

5.1 Benefits of Appropriate Task Routing and Assignment in Crowdsourcing

There are a number of benefits to crowdsourcing research and practice that would result from more appropriate task routing and assignment. Here, we explore one way of achieving this, which entails using a person's cognitive abilities to predict their performance on crowdsourcing tasks. The strong correlation between the average performance of the cognitive tests and crowdsourcing tasks (Figure 2), the similarity between the results of both our models (Table 2) and the improvements shown after ordinary bootstrapping (Table 3) provide evidence towards this assertion and therefore warrants additional exploration within this research agenda. Ultimately, creating reliable, even if not-optimal, predictions about workers' performance based on elicitation of their cognitive abilities may pave the way to efficient crowdsourcing task routing and assignment. This would allow crowdsourcing platforms and/or task requesters to find more suitable workers to be assigned to a proposed task, which in turn is likely to have an effect on the quality of the crowdsourced data.

Another potential benefit of improvements to task routing and assignment, is that it might mitigate the need for error controlling approaches that are currently predominantly used in crowdsourcing, such as the Gold Standard [10]. This can be quite important in certain situations, as authoring gold data can be burdensome, and gold standards are challenging to implement in subjective or generative tasks, like writing an essay [31]. Quality assurance mechanisms that use worker agreement [4,17,28] as the quality metric can also benefit from better task routing and assignment. The likelihood of collecting tasks with high disagreement between workers is lower, and therefore less data will potentially be discarded. This inherently leads to an increase in reliability of the collected data.

Furthermore, previous work has established that task difficulty has a significant effect on task performance [47]. Specifically, for tasks with higher difficulty levels, workers are more likely to give up, or provide an approximate or even non-serious answer for the task. While replicating classic experiments in an online crowdsourcing market, Horton *et al.* [25] also found a relationship between task difficulty and likelihood to complete tasks. However, what a worker considers difficult can in some cases be a direct result of their own cognitive abilities. This was confirmed during our interviews, as participants reported different tasks as being the most challenging to them. By nudging workers to complete tasks that they are more likely capable of completing well can lead to higher task uptake and reduced likelihood of desultory responses.

5.2 Measuring Cognitive Abilities on Crowdsourcing Platforms

While we provide evidence of the relationship between cognitive abilities and crowdsourcing performance, gathering this information about workers can be challenging. One way to achieve this in online crowdsourcing platforms would be to have workers complete cognitive abilities assessment tests. This would be particularly useful when considering new workers of a platform, as there is no past performance to predict how well they will do on similar or relevant tasks, so collecting this information upon or shortly after registration would be ideal. Further, in some cases, these new workers without prior established performance may pursue tasks that provide them the best payout instead of tasks that they are more suited to complete. In literature, person-job misfit has been shown to substantially affect performance and place increasing strain on both sides over time [6]. Further, in our interviews participants highlighted the importance of feeling capable of achieving their tasks and how the absence of this sentiment could affect their performance. Hence, it is in the platform's best interest to appropriately route workers to suitable tasks by, for example, providing a list of suggested tasks based upon their measured cognitive abilities.

Our findings also have implications for the situated crowdsourcing research agenda. Situated crowdsourcing entails embedding input mechanisms (*e.g.*, public displays, tablets) in a physical space and leverage users' serendipitous availability [39] or idle time ("cognitive surplus" [50]). While situated crowdsourcing may be better suited for "local" tasks [27], it has also been shown to be effective with typical crowdsourcing tasks that can be seen in online labour markets [26]. While some situated crowdsourcing deployments do not track workers, making it impossible to assign task based on an individual's cognitive abilities [18,20,27], others have tracked individual workers, such as Bazaar, a situated crowdsourcing market that had user accounts, a virtual currency and rewards [26]. Here, as with online crowdsourcing platforms, having an initial cognitive assessment stage could be beneficial. A potential issue in this scenario is that workers may not be willing to undertake a long pre-requisite cognitive abilities test before being able to start completing crowdsourcing tasks due to the increased barrier for participation, which can significantly hinder uptake on situated crowdsourcing [16,18]. With that in mind, simplified versions of these tests could be constructed and validated, or workers could be financially rewarded for taking those tests. While such simplified versions of cognitive tests may be less reliable than a thorough elicitation of cognitive abilities, we argue that even minor indications of workers' cognitive abilities can improve the quality of workers' contributions, or their uptake of work.

5.3 Limitations

We acknowledge several limitations in the presented study. First, the sample size used in the study is relatively small. We took steps to mitigate this limitation when creating our prediction models by performing cross-validation (LOO) and running ordinary bootstrapping, a commonly used and robust technique when dealing with relatively small sample sizes. However, to fully establish our findings, further research with larger samples sizes should be considered. Second, while visual and fluency based tasks are commonplace in crowdsourcing platforms, we did not investigate other types such as those that require only rational thinking. Finally, we conducted a controlled lab study instead of recruiting actual workers from an online crowdsourcing platform. It was important for us to appropriately monitor the data collection, which allowed us to follow precisely the instructions set by the ETS Cognitive Tests and also avoid other issues such as collusion or worker distraction.

6. CONCLUSION AND FUTURE WORK

In this paper, we tackle a fundamental challenge in crowdsourcing: how to appropriately route and assign workers to suitable crowd-tasks. Here we highlight the role that workers' cognitive abilities have on crowdsourcing performance, specifically when considering visual and fluency-based tasks. Our results indicate that by measuring workers' cognitive abilities it is possible to reliably predict their crowdsourcing performance. Our findings have substantial implications to the crowdsourcing research agenda by providing a new mechanism for *a priori* worker allocation. In future work we will extend our analysis to include a wider variety of crowdsourcing task types and a larger number of participants. We also intend to develop and validate a set of crowdsourcing tests that are able to measure workers' cognitive abilities. By using these tests, researchers and task requesters would not have to rely on tools such as the ETS Kit of Factor-Referenced Cognitive Tests, which can be costly, time-consuming and may not reliably capture these abilities in an online scenario.

7. ACKNOWLEDGMENTS

This work is partially funded by the Academy of Finland (Grants 276786-AWARE, 286386-CPDSS, 285459-iSCIENCE, 304925-CARE), and the European Commission (Grants PCIG11-GA-2012-322138, 645706-GRAGE, and 6AIKA-A71143-AKAI). We thank all participants that took part in our experiment.

8. REFERENCES

[1] Allen, G. L., Kirasic, K. C., Dobson, S. H., Long, R. G., and Beck, S. 1996. Predicting environmental learning from spatial abilities: An indirect route. *Intelligence* 22, 3: 327-355.

[2] Bandura, A. 2001. Social cognitive theory: An agentic perspective. *Annual Review of Psychology* 52, 1: 1-26.

[3] Bernstein, A., Klein, M., and Malone, T. W. 2012. Programming the global brain. *Communications of the ACM* 55, 5: 41-43.

[4] Callison-Burch, C. 2009. Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 286-295.

[5] Caplan, R. D. 1987. Person-environment fit theory and organizations: Commensurate dimensions, time perspectives, and mechanisms. *Journal of Vocational Behavior* 31, 3: 248-267.

[6] Chilton, M. A., Hardgrave, B. C., and Armstrong, D. J. 2005. Person-job cognitive style fit for software developers: the effect on strain and performance. *Journal of Management Information Systems* 22, 2: 193-226.

[7] Chiu, C., Hsu, M., and Wang, E. T. 2006. Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories. *Decision Support Systems* 42, 3: 1872-1888.

[8] Donkor, B. 2013. On Social Sentiment and Sentiment Analysis. Retrieved 13/05/2016 from http://brnrd.me/social-sentiment-sentiment-analysis/

[9] Downing, R. E., Moore, J. L., and Brown, S. W. 2005. The effects and interaction of spatial visualization and domain expertise on information seeking. *Computers in Human Behavior* 21, 2: 195-209.

[10] Downs, J. S., Holbrook, M. B., Sheng, S., and Cranor, L. F. 2010. Are Your Participants Gaming the System?: Screening Mechanical Turk Workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2399-2402. http://dx.doi.org/10.1145/1753326.1753688

[11] Dunnette, M. D. 1976. Aptitudes, abilities, and skills. *Handbook of Industrial and Organizational Psychology*: 473-520.

[12] Efron, B. 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* 78, 382: 316-331.

[13] Ekstrom, R. B., French, J. W., Harman, H., and Dermen, D. 1976. *Manual for kit of factor referenced cognitive tests.* Educational Testing Service Princeton, NJ.

[14] Feldman, M., and Bernstein, A. 2014. Cognition-based Task Routing: Towards Highly-Effective Task-Assignments in Crowdsourcing Settings.

[15] Ferrari, S., and Cribari-Neto, F. 2004. Beta regression for modelling rates and proportions. *Journal of Applied Statistics* 31, 7: 799-815.

[16] Goncalves, J., Ferreira, D., Hosio, S., Liu, Y., Rogstadius, J., Kukka, H., and Kostakos, V. 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*, ACM, 753-762. http://dx.doi.org/10.1145/2493432.2493481

[17] Goncalves, J., Hosio, S., Ferreira, D., and Kostakos, V. 2014. Game of Words: Tagging Places through Crowdsourcing on Public Displays. In *Proceedings of the 2014 Conference on Designing Interactive Systems*, 705-714. http://dx.doi.org/10.1145/2598510.2598514

[18] Goncalves, J., Hosio, S., Rogstadius, J., Karapanos, E., and Kostakos, V. 2015. Motivating participation and improving quality of contribution in ubiquitous crowdsourcing. *Computer Networks* 90, 34-48.

[19] Goncalves, J., Kukka, H., Sánchez, I., and Kostakos, V. 2016. Crowdsourcing Queue Estimations in Situ. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing,* ACM, 1040-1051. http://dx.doi.org/10.1145/2493432.2493481

[20] Goncalves, J., Pandab, P., Ferreira, D., Ghahramani, M., Zhao, G., and Kostakos, V. 2014. Projective testing of diurnal collective emotion. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM, 487-497. http://dx.doi.org/10.1145/2818048.2819997.

[21] Ho, C., Jabbari, S., and Vaughan, J. W. 2013. Adaptive task assignment for crowdsourced classification. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 534-542.

[22] Ho, C., and Vaughan, J. W. 2012. Online Task Assignment in Crowdsourcing Markets. In *AAAI Conference on Artificial Intelligence*, 12, 45-51.

[23] Hoffman, B. J., and Woehr, D. J. 2006. A quantitative review of the relationship between person--organization fit and behavioral outcomes. *Journal of Vocational Behavior* 68, 3: 389-399.

[24] Horowitz, D., and Kamvar, S. D. 2010. The anatomy of a large-scale social search engine. In *Proceedings of the 19th International Conference on World Wide Web*, 431-440.

[25] Horton, J., Rand, D. G., and Zeckhauser, R. J. 2011. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics* 14, 3: 399-425. http://dx.doi.org/10.1007/s10683-011-9273-9

[26] Hosio, S., Goncalves, J., Lehdonvirta, V., Ferreira, D., and Kostakos, V. 2014. Situated Crowdsourcing Using a Market Model. In *User Interface Software and Technology*, ACM, 55-64. http://dx.doi.org/10.1145/2642918.2647362

[27] Hosio, S., Goncalves, J., Kostakos, V., and Riekki, J. 2015. Crowdsourcing public opinion using urban pervasive technologies: Lessons from real-life experiments in Oulu. *Policy & Internet* 7, 2: 203-222.

[28] Ipeirotis, P. G., Provost, F., and Wang, J. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on Human Computation*, 64-67.

[29] Jung, H. J. 2014. Quality assurance in crowdsourcing via matrix factorization based task routing. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, 3-8.

[30] Kanfer, R., and Ackerman, P. 1989. Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology* 74, 4: 657.

[31] Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., and Horton, J. 2013. The future of crowd work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, ACM, 1301-1318. http://dx.doi.org/10.1145/2441776.2441923

[32] Kristof, A. L. 1996. Person-organization fit: An integrative review of its conceptualizations, measurement, and implications. *Personnel Psychology* 49, 1: 1-49.

[33] Kristof-Brown, A.L., Zimmerman, R. D., and Johnson, E. C. 2005. 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.

[34] Lee, M. D., Steyvers, M., De Young, M., and Miller, B. 2012. Inferring expertise in knowledge and prediction ranking tasks. *Topics in Cognitive Science* 4, 1: 151-163.

[35] Li, H., Li, T., and Wang, Y. 2015. Dynamic Participant Recruitment of Mobile Crowd Sensing for Heterogeneous Sensing Tasks. In *Mobile Ad Hoc and Sensor Systems (MASS), 2015 IEEE 12th International Conference on*, 136-144.

[36] Mao, A., Kamar, E., and Horvitz, E. 2013. Why stop now? predicting worker engagement in online crowdsourcing. In

First AAAI Conference on Human Computation and Crowdsourcing.

[37] Mayer, R. E. 2014. Cognitive theory of multimedia learning. *The Cambridge Handbook of Multimedia Learning*: 43-71.

[38] Morris, R., Dontcheva, M., and Gerber, E. M. 2012. Priming for better performance in microtask crowdsourcing environments. *Internet Computing, IEEE* 16, 5: 13-19.

[39] Müller, J., Alt, F., Michelis, D., and Schmidt, A.. 2010. Requirements and design space for interactive public displays. In *Proceedings of the international conference on Multimedia*, ACM, 1285-1294. http://dx.doi.org/10.1145/1873951.1874203

[40] Navarro, G. 2001. A Guided Tour to Approximate String Matching. *ACM Comput. Surv.* 33, 1: 31-88. http://dx.doi.org/10.1145/375360.375365

[41] Peters, T. A. 1996. Human factors in information systems: Emerging theoretical bases. *Human factors in information systems: Emerging theoretical bases*. http://dx.doi.org/10.1002/(SICI)1097-4571(199608)47:8.

[42] Pinker, S. 1984. Visual cognition: An introduction. *Cognition* 18, 1: 1-63.

[43] Plass, J. L., Moreno, R., and Brünken, R. 2010. *Cognitive load theory.* Cambridge University Press.

[44] Reber, R., and Schwarz, N. 1999. Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition* 8, 3: 338-342.

[45] Reddy, S., Estrin, D., and Srivastava, M. 2010. Recruitment Framework for Participatory Sensing Data Collections. In *Proceedings of the 8th International Conference on Pervasive Computing*, Springer-Verlag, 138-155. http://dx.doi.org/10.1007/978-3-642-12654-3_9

[46] Rodrigues, A. 2014. The theory of cognitive dissonance: a current perspective. *Arquivos Brasileiros de Psicologia Aplicada* 22, 2: 126-127.

[47] Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., and Vukovic, M. 2011. An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. In *International AAAI Conference on Web and Social Media*, 321-328.

[48] Ruble, T. L., and Cosier, R. A. 1990. Effects of cognitive styles and decision setting on performance. *Organizational Behavior and Human Decision Processes* 46, 2: 283-295.

[49] Shirani-Mehr, H., Banaei-Kashani, E., and Shahabi. C. 2009. Efficient Viewpoint Assignment for Urban Texture Documentation. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ACM, 62-71. http://dx.doi.org/10.1145/1653771.1653783

[50] Shirky, C. 2010. *Cognitive surplus: How technology makes consumers into collaborators.* Penguin.

[51] Speier, C., Valacich, J. S., and Vessey, I. 1999. The Influence of Task Interruption on Individual Decision Making: An Information Overload Perspective. *Decision Sciences* 30, 2: 337-360. http://dx.doi.org/10.1111/j.1540-5915.1999.tb01613.x

[52] Tan, F. B., and Hunter, M. G. 2002. The repertory grid technique: A method for the study of cognition in information systems. *MIS Quarterly*: 39-57.

[53] Vakharia, D., and Lease, M. 2013. Beyond AMT: An analysis of crowd work platforms. *arXiv preprint arXiv:1310.1672*.

[54] Van Belle, G. 2011. *Statistical rules of thumb.* John Wiley & Sons.

[55] Van Merrienboer, J., and Sweller, J. 2005. Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review* 17, 2: 147-177.

[56] Verquer, M. L., Beehr, T. A., and Wagner, S. H. 2003. A meta-analysis of relations between person--organization fit and work attitudes. *Journal of Vocational Behavior* 63, 3: 473-489.

[57] Williams, M. N., Grajales, C. A., and Kurkiewicz, D. 2013. Assumptions of multiple regression: correcting two misconceptions. *Practical Assessment, Research & Evaluation* 18, 11: 2.

[58] Wilson, R. S., De Leon, C. F., Barnes, L., Schneider, J. A., Bienias, J. L., Evans, D. A., and Bennett, D. A. 2002. Participation in cognitively stimulating activities and risk of incident Alzheimer disease. *Jama* 287, 6: 742-748.

[59] Xiao, M., Wu, J., Huang, L., Wang, Y., and Liu, C. 2015. Multi-task assignment for crowdsensing in mobile social networks. In *Computer Communications (INFOCOM), 2015 IEEE Conference on*, 2227-2235.

[60] Zhang, H., Horvitz, E., Chen, Y., and Parkes, D. C. 2012. Task routing for prediction tasks. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, 889-896.