Cognition-based Task Routing:Towards Highly-Effective Task-Assignments in Crowdsourcing Settings

Feldman, Michael; Bernstein, Abraham

Abstract: In recent years the rising popularity of outsourcing work to crowds has led to increasing importance to find an effective assignment of suitable workers with tasks. Even though attempts have been made in related areas such as expertise identification most crowdsocuring jobs today are assigned without any predefined policy. Whilst some have investigated assigning jobs based on availability or experience no dominant method has been identified so far. We propose an assignment of tasks to crowd-workers based on their cognitive capability, by conducting a set of cognitive tests and comparing them with performance on typical crowdtasks. Moreover, we examine different setups to predict task performance where a) cognitive abilities, b) performance on previous crowdtasks, or c) both of them, are partially known. Preliminary results show that cognition-based task assignment leads to an improvement in task performance prediction and may pave the way to more intelligent crowd-worker recruitment.

Cognition-based Task Routing: Towards Highly-Effective Task-Assignments in Crowdsourcing Settings

Research-in-Progress

Michael Feldman
Department of Informatics
University of Zurich
Binzmühlestrasse 14, 8050 Zürich
Switzerland
feldman@ifi.uzh.ch

Abraham Bernstein
Department of Informatics
University of Zurich
Binzmühlestrasse 14, 8050 Zürich
Switzerland
bernstein@ifi.uzh.ch

Abstract

In recent years the rising popularity of outsourcing work to crowds has led to increasing importance to find an effective assignment of suitable workers with tasks. Even though attempts have been made in related areas such as expertise identification most crowdsocuring jobs today are assigned without any predefined policy. Whilst some have investigated assigning jobs based on availability or experience no dominant method has been identified so far. We propose an assignment of tasks to crowd-workers based on their cognitive capability, by conducting a set of cognitive tests and comparing them with performance on typical crowdtasks. Moreover, we examine different setups to predict task performance where a) cognitive abilities, b) performance on previous crowdtasks, or c) both of them, are partially known. Preliminary results show that cognition-based task assignment leads to an improvement in task performance prediction and may pave the way to more intelligent crowd-worker recruitment.

Keywords: Crowdsourcing, Task assignment, Cognition/cognitive science

Introduction

Crowdsourcing has gained increased relevance as an accepted approach for outsourcing activities to an online community. In recent years, the business community embraced the approach of outsourcing some activities to a crowd by means of evolved specialized, web-based platforms. Jobs are mostly partitioned into a group of simplified sub-tasks and distributed to crowd-workers in an open call manner. Encouraged by human computation potentials, attempts have been made recently to extend the types of human computation tasks beyond relatively simple and non-demanding ones. As a consequence, the crowdsourcing domain is faced with an emerging need to find concepts and paradigms such that complex tasks, requiring wide spectrum of human abilities and talents, can be successfully assigned to the suitable crowd-workers. Finding appropriate crowd-workers, however, is non-trivial due to the human motivational, cognitive, and error diversity (Bernstein et al. 2012). Moreover, the remote and unstable character of most crowd markets, where the ability to track and profile workers is limited, gives rise to an even greater challenge. Many have explored the motivational and error diversity in the crowdsourcing contexts. Masson and Watts (2010), for example, point out that higher financial reward does not necessarily lead to improved work quality as the "anchoring effect" has an impact on workers' perceived value of the provided output. Therefore, efforts were investigated to establish effective incentives mechanisms to promote performance-based reward (e.g., Minder et al. 2012). Substantial work has been done to find error control policies for quality control (e.g., Bernstein et al. 2010). To date, most common quality control policies focus on controlling error via "smart" result aggregation following the elicitation of multiple answers for the tasks. To avoid interdependencies, some also split the crowd into two groups: one that solves the task and a second that estimates result quality (Bernstein et al. 2012). Another common mechanism based on ground truth embedding into the workflow, such that worker's creditability is estimated based on her responses to those questions (Oleson et al. 2011). These approaches assume that the cognitive diversity, the capability of the crowd-workers assigned to a given task, cannot be controlled.

This paper focuses on finding the means to address cognitive diversity (i.e., finding the person whose cognitive capabilities are best suited to the requirements of a given task). In our empirical exploration, we concentrate on visual tasks, as they are often published on crowd platforms and require wide range of implicit abilities. In addition, they are susceptible to a high degree of cognitive variance (Pinker 1984). Furthermore, visual tasks are influenced by norms or cultural preferences (e.g., Reinecke and Bernstein 2013) furthering the need for an appropriate task assignment. Ignoring the diversity of workers' cognitive abilities will almost certainly lead to a mismatch in task-worker assignment, and therefore result in inferior performance. Our focus on visual tasks may limit the generalizability of our results, as the generalization from visual to other cognitive tasks would first have to be established. We explore if a crowd-worker's cognitive profile can predict the quality of her work. To elicit a crowd-workers cognitive capabilities we use the well-established *factor-referenced cognitive tests kit* introduced by Ekstrom et al. (1976) and provided by the Educational Testing Service (ETS). We measure the cognitive abilities as well as performance on visual crowd-tasks of participants and conduct analysis to determine whether it is possible to predict (i) a crowd-worker's performance in crowd-tasks based on her cognitive profile, (ii) a crowd-worker's cognitive parameters based on a limited set of performed crowd-tasks, and (iii) crowd-task performance based on similar, previously performed tasks.

There are at least two potential contributions of this ongoing research. First, creating reliable predictions about workers' performance based on elicitation of either their cognitive qualities or prior performance on their tasks may pave the way to efficient crowd-job assignment. This will allow finding suitable workers for a proposed job and might mitigate the need for error controlling approaches currently so dominant in crowdsoucring. Second, an analysis of the proposed cognitive approach in conjunction with other influencing factors such as declared skills, personal details (e.g., age, gender, or geography), availability, incentive mechanisms, and reputation could lead to a better understanding of the interrelationships between all factors underlying the human performance on crowd platforms. This will enable to investigate general approaches for tasks-assignment and could help the recruitment of crowd-workers in optimal manner. As our research is ongoing, we intend to explore all of above mentioned research directions as a part of further research.

Next, we succinctly review the relevant literature setting the stage for an introduction of our research hypotheses. We continue by presenting our experimental design, the analysis approach, and preliminary results before we close with a discussion of limitations and some future work.

Literature Review

Numerous studies conducted in industrial and organizational psychology have recognized cognitive abilities as a substantial factor for determining the work performance of individuals (e.g., Kanfer and Ackerman 1989). These theories address the exploration of the relationship between cognitive characteristics and task performance, highlighting the importance of cognitive abilities in predicting individual differences in job performance. Moreover, the findings are consistently point out that high cognitive capabilities of workers lead to better job performance (Dunnette 1976; Hunter 1986). As human behavior may be represented as a mixture of personal factors, behavior presets, and the social impact modifications of this environment may have critical affect on human functioning, and play a key role in human based systems (Chiu et al. 2006). Some of the cognitive theories as *Social Cognitive Theory* (Bandura 2001)*, Personal Construct Theory* (Tan and Hunter 2002), *Cognitive Load Theory (*Sweller et al. 2011*), or Cognitive Dissonance Theory (*Rodrigues 2014*)* have been widely applied in the information systems (IS) research, thereby gaining increasing credibility. The adoption of these theories was driven by growing understanding that human behavior needs to be reviewed as a conglomerate of cognitive and social processes. Studies based on these approaches can be found in IS domains such as Human Computer Interaction (HCI), computer systems design, information systems adoption and information presentation *(*Allen 1996; Peters 1996).

In last decade, the IS research field witnessed a substantial change as rapid development of collaborative systems like Wikipedia, Linux, Yahoo! Answers, and Amazon's Mechanical Turk that incorporate multitude of humans and help solve a wide variety of problems. The problems vary from simple, as reviewing products, to complex, emerging systems that intend to store large-scale human knowledge with supporting interrelationship structure of represented artifacts (e.g., Wikipedia, StackOverflow) (Doan et al 2011) and are often referred to as Crowdsourcing. Crowdsourcing systems are usually web-based collaborative systems that support online tasks performed by distributed crowd-workers, which may or may not be compensated financially. Hence, crowd-work is a sociotechnical work system constituted through a set of relationships that connect organizations, individuals, technologies, and work activities (Kittur et al. 2013). Due to the world-wide distributed multitude of the workers and the variety of their characteristics, crowdsourcing often seeks to solve cognitively complex, large-scale tasks by decomposing them into micro-tasks, which are then executed by crowd-workers in a distributed fashion. Specifically, a requester posts a problem or a set of decomposed micro-problems online, a vast number of individuals deliver solutions (and may receive in return some reward such money, fame, or just the feeling of contributing), the results are then harvested and aggregated, such that the requester's benefit is maximized (Lee et al. 2012). Crowdsourcing has gained popularity in multiple domains such as in science, marketing, geospatial processing, disaster recovery, military, or software development. The military is exploring ways to collect intelligence data through crowdsourcing, government agencies are using it to collect data on everything from road repairs to urban planning, and relief agencies are turning to it to better understand how to focus aid and resources. Crowdsourcing-driven companies enable people throughout the world to map everything from natural disasters to political turmoil. Marketing companies implement crowdsourcing in sake of product development, advertising and promotion, and marketing research. Geographic information created by amateur citizens used for mapping and companies implement crowdsourced information about updated traffic to ease traffic and avoid traffic jams (Whitla 2009; Goodchild and Glennon 2010; Greengard 2011).

One popular crowdsourcing platform, Mechanical Turk, was introduced in 2005 by Amazon as an environment for humans to perform tasks that are very difficult or impossible for computers, such as extracting data from images, image labeling, audio transcription, classification of content, and massive translation. Thus, the platform is a labor market for micro-tasks, where every task is a tiny fraction of the proposed problem and the typical contribution can be accomplished in minutes (Mason and Suri 2011). Whilst the time and cognitive effort for every single task is small the combination of the micro-tasks can result in major accomplishments. However, viewing crowd-workers as computational units ignores the underlying mechanisms of cognition as complex emotions, creativity, and high-order thinking. Whilst simple tasks are currently the most common ones in crowd-sourcing marketplaces, creative crowd-jobs are expected to gain complexity and become more prevalent over the years to come (Morris et al. 2012, 2013). In fact, a variety of crowd platforms as oDesk, Elance or CrowdSource promote tasks that require high level of expertise and diverse talents (Vakharia and Lease, 2013). Hence, the cognitive aspect

increasingly plays a key role in workers performances variability and its management may improve the quality of the work significantly.

The cognitive science community has been active in *expertise assessment* research, exploring cognitive and other aspects of intelligent expert-job matchmaking. Lee et al. (2012) developed cognitive models for measurement of expertise using the differences between responses of workers. The approach is limited by the dependency on large number of workers, required for the effective assessment via cross-examination of responses to the tasks. Attempts have been done to fit a mathematical model on observed data of individuals. For instance, Weiss et al. (2003, 2014) developed the *Cochran-Weiss-Shanteau* (CWS) index to distinguish between experts and non-experts where consistency, validity, and discrimination are the underlying characteristics of expertise. While validity requires ground truth and is usually difficult to establish, the other two properties are readily observable, and are combined in the CWS index. Mao et al. (2013) design and construct statistical models that provide predictions about the forthcoming engagement of volunteers with respect to different sets of features that describe user behavior. Sculley et al. (2009) explores supervised learning for making predictions about individual behavior on the web. Zhang et al. (2007) analyzed a large online help-seeking community, using social network analysis methods, to identify users with high expertise algorithms. The testing of network-based ranking algorithms revealed that they did nearly as well as human raters. However, there were significant tradeoffs among the algorithms.

The theoretical foundations of our study may be seen in the Person-Job fit element of the well-studied Person-Environment fit theory. Person-job fit was defined as the fit between the abilities of a person and the demands of a job or the desires of a person and the attributes of a job (Kristof 1996). Specifically, the effect of cognitive abilities on performance has been broadly discussed in series of studies (Verquer et al. 2003; Hoffman and Woehr 2006; Hunter 1986). Chilton et al. (2005), in particular, examine the impact of Person-Job fit on performance and strain. They show that high level of person-job cognitive style misfit affects performance and strain. However, it is noteworthy that other studies did not find a significant link between cognitive abilities and performance (Kristof-Brown et al. 2005; Ruble and Cosier 1990). As far as we know our study is the first attempt to establish this connection in the crowdsourcing domain and exploit it for task assignment purposes.

In the *crowdsourcing research community*, Zhang et al. (2012) introduce principles and methods for task routing that aim to harness people's abilities to jointly contribute to a task and to route tasks to others who can provide further contributions. Jung (2014) describes a method to predict a crowd worker's accuracy on new tasks based on his accuracy on past tasks by means of collaborative filtering. The underlying idea is to model similarity of past tasks to the target task such that past task accuracies can be optimally integrated to predict target task accuracy. Horowitz and Kamvar (2010) are using social network to route the task to suitable workers. 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 an assessment of who is best able and willing to provide an answer.

These studies provide first insight on how to assign tasks to workers. None of them, however, offers a dominating strategy for doing this assignment with a consistent performance. It is the underlying assumption of this paper that the cognitive capability of a crowd-worker needs to take a prime contribution in such an assessment. For measuring cognitive abilities in our experiments we have chosen to use Factor-Referenced Cognitive Tests constructed by the Educational Testing Service (Ekstrom et al. 1976). The tests consist 72 factor-referenced cognitive tests for 23 factors and aim to serve as a measurement for cognition dimensions. The kit of tests was published in 1976 and has gained validity and reliability across disciplines and over time. The tests were implemented in various domains such as multimedia learning, Alzheimer's disease research, decision-making, or human spatial cognition (Wilson et al. 2002; Mayer 2005; Speier et al. 1999; Allen et al. 1996). As our experiments rely on visual and spatial cognitive abilities it is notable that many researches highlight the appropriateness of ETS cognitive tests as a measure to cognition ability abstract in this field (Downing et al. 2005; Velez et al. 2005).

Research Hypotheses

In this section we present our research questions that focus on the interrelations between visual crowd-tasks performance and cognitive capabilities. The research questions and corresponding hypothesis are arranged to increasingly investigate the dependency between crowd-workers' cognitive capabilities and tasks performance. Starting with exploring the predictive power of past tasks' performances over future

tasks, we then extend the discussion by considering the cognitive aspect as a substantial underlying factor to predict a crowd-worker's work quality.

Research Question 1 (RQ1): Can knowledge about past task performance be used to predict future task performance?

The question addresses the setup where crowd-workers' previous task performance is known and evaluated. Therefore, this prior knowledge may be used to predict future performance from two perspectives. First, an individual's outputs quality may be consistent for the identical task conducted in repeatable manner over time (see H1a). Second, relying on the assumption that performances on similar tasks are significantly correlated, future performance on similar tasks might be predicted (see H1b). In both cases the analysis is based on crowdsourced accomplishments, regardless with underlying factors that may lead to performance diversity. This research question establishes both a control and a baseline to improve over for our study.

Hypothesis 1a (H1a): A crowd-worker's performance on a given task predicts future performance on the same task (task performance consistency).

As crowdsourcing markets as of today mostly apply to limited pool of relatively simple and monotonous tasks such as labeling pictures, classifying content, or transcription it is reasonable to expect worker to work on the same job over and over again along the time. Therefore, we look into consistency of performance on the same kind of tasks controlling for all effects but learning and fatigue.

H1b: A crowd-worker's performance on some previous tasks predicts future performance on another task.

This second hypothesis examines whether similar tasks may be helpful in predicting future task with similar requirements. In context of this research all designed tasks belong to visual domain. It is, therefore, reasonable to assume a great overlap in required skills, capabilities, and other actuating factors. This hypothesis is examined by means of established, off-the-shelf collaborative filtering techniques as well as data imputation methods. This hypothesis establishes a baseline for prediction.

RQ2: Is there is a connection between test results on standardized visual cognitive tests and task performance in crowd-tasks?

The question deals with the interplay between cognitive abilities and crowd-tasks results. Therefore, in contrast to the first research question, the cognitive capabilities of workers are considered. This research question reduces the traditional view of *Person-Job fit* into a setting where cognitive abilities are assumed to be direct predictors of job performance and transfers it to the crowdsourcing area. The reduction is founded in cognitive load theory (and its "descendant" theories), whereby we assume that a person's cognitive abilities are a good predictor of his/her general abilities. First, the correlation between cognitive tests and crowd-tasks performances is examined (H2a). Then, the directionality is being hypothesized (H2b). Finally, the predictive power of cognitive abilities over crowd-tasks performance is examined (H2c). We operationalize this question via the following hypotheses:

H2a: A crowdworker's performance on standardized visual cognitive tests correlates with her performance on visual crowd-tasks.

H2b: The better the visual cognitive abilities of crowd-workers, the better they perform on visual crowd-tasks.

H2c: A crowdworker's performance on the standardized cognitive tests (such as the ETS cognitive test) predicts future performance on another typical crowdsourcing task.

Having established a possible connection between standardized tests and task performance the last RQ addresses the question whether the inclusion of cognitive factors improves prediction performance over methods that do not consider this dimension. Specifically, it states:

RQ3: Do standardized cognitive tests improve prediction quality for crowd tasks over non-cognition based methods?

This research question compares proposed approach with existing approaches, such as averaged rating or collaborative filtering.

H3a: A crowdworker's performance prediction based on standardized cognitive tests outperforms a method based on averaged rating, data imputation, and collaborative filtering.

H3b: A (learned) mapping between crowdworkers' performance on standardized cognitive tests and crowdsourced tasks can be used to predict a new crowdworker's performance on a crowdsourcing tasks, who has not taken the cognitive tests.

Experimental Design and Data Collection

Our experiment included six ETS cognitive tests measuring cognitive qualities in visual tasks as well as five visual crowd-tasks. The ETS tests included the Hidden Patterns, Hidden Figures, Card Rotation, Cube Comparison, Paper Folding, and the Surface Development 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 the tests (Ekstrom et al. 1976). The crowdsourcing tasks where designed inspired by typical tasks posted on MTurk. They included (i) a *text distortion* task, where subjects were asked to restore visually distorted sentences akin to a captcha, (ii) a *distance evaluation* task, where subjects were asked to evaluate which of two objects is closer, (iii) a *height evaluation* task, where subjects were asked to evaluate which of two objects is higher, (iv) an *item recognition* task, where subjects have to ascertain if certain items are shown in a picture, and (v) a *item classification task*, were subjects were asked to classify depicted items into one of the four classes. All tasks included questions that varied in their complexity and were designed to cover different aspects of visual perception. To avoid learning effects the tasks were shuffled.

We believe that the workers' cognitive abilities become more important in conducting time-consuming, complex tasks (rather than short term micro-tasks). We, therefore, chose the freelance platform Elance to conduct the experiment. Elance is a crowdsourcing platform that contains relatively complex jobs (e.g., content translation, design) and promotes long-term contacts between requesters and workers. It has a large number of diverse workers (i.e., freelancers) distributed around the world.

Data collection was performed in two steps. First, a pilot experiment included 11 subjects with 7 cognitive tests and 5 crowd-tasks. As a result, one non-informative cognitive test was excluded and the crowd-tasks were modified to reduce confusion. Second, we ran the experiment by recruiting 37 (new) subjects, mostly from English speaking countries, with preliminary short on-line chats to reduce possible misunderstandings of the instructions. To compensate the subjects for their effort of participating in the roughly 80 minutes experiment there were paid 15 USD and positive feedback, based on their performances (i.e., depending on performance percentile).

Data Analysis and Preliminary Results

To evaluate the quality of our predictions for all hypotheses we use two types of performance metrics. First, to evaluate the pure prediction quality we employ accuracy, root mean squared error (RMSE), and mean absolute error (MEA). Second, we explore how good the prediction is as a decision factor to choose suitable crowd-workers. For this we do not need to check the prediction quality of all crowd-workers, as one typically only assigns tasks to top-rated workers. Hence, we will use margin curves such as precision, recall, or the area under ROC curves. Given that this paper is a research in progress submission we will only show that preliminary results for RMSE, MEA, and correlation, where appropriate. The margin curve analysis will follow in the full publication.

We tested the **H1a** on the item classification task containing 25 similar questions. Essentially, we regarded the series of 25 classifications as a sequence of tasks, where the first n classifications predict the performance on the $n+1^{th}$ classification. This parallels the performance evaluation of existing crowdsourcing platforms, where workers' rank is calculated as an averaged feedback of employers for previously performed highly similar tasks. A preliminary analysis can be found in Figure 1, which shows a boxplot summarizing the results for all 37 subjects. The Figure clearly shows that with the exception of one outlier the performance improves and the variance decreases with an increased number of questions answered. We intend to extend the analysis for the other tasks and compute the performance metrics.

In our preliminary analysis we examined ***H1b*** by means of the Singular Value Decomposition (SVD) approximation for collaborative filtering (Zhang et al. 2005) and Multivariate Imputations by Chained

Equations (MICE) (White and al. 2011). Both methods are often used to predict missing values in multivariate datasets. As a preliminary analysis we randomly omited 30% of the subjects' performances in the crowdtasks and estimated the missing values. The results reveal that SVD outperforms MICE both in terms of MAE (10.2% versus 11.8%) and RMSE (12.7% versus 15.6%). As witnessed by these results we have reason to believe that H1b can be supported by the data. For our final analysis we will both employ further approaches as well as test the sensitivity of these results towards various percentages of missing values/deletions.

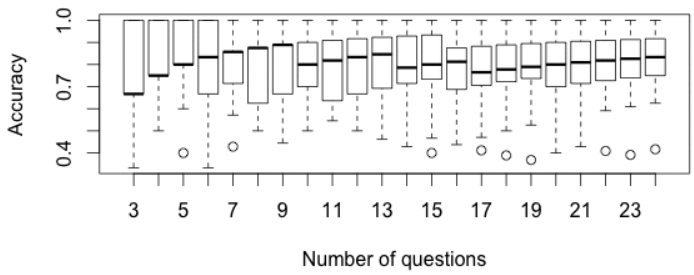

Figure 1. Improving prediction accuracy based of prior performance throughout the 25 questions

Hypotheses ***H2a*** and ***H2b*** address the correlation between the averaged performance on cognitive tests and averaged performance on visual crowd-tasks. Whilst our data is not normally distributed the Wilk-Shapiro test is close to significant for most of the response variables. We, therefore, report both the Pearson product-moment 0.664 and the Spearman correlation coefficient ρ_s = 0.644 (significant at p<<0.001). This gives us reason to believe that there is a somewhat strong positive correlation. We will explore these hypotheses further by investigating the robustness of these results against outliers and by drawing new samples.

We explored ***H2c*** by building predictive models for crowd-tasks performances, using the cognitive tests' performances as explanatory variables. We built three predictive models: linear and second order polynomial (PR2) regressions and a generalized linear model (GLM). We omitted a linear regression as our data is not normally distributed (as tested by Wilk-Shapiro test for normality) and the performance of the linear model is inferior to PR2. As for the GLM, the currently collected data fully meets the underlying assumptions of the model (Breslow, 1996). Table 1 provides with results of two models in terms of MAE, RMSE and correlation of predicted vs. observed performances over crowd tasks.

		Text distortion	Distance evaluation	Height evaluation	Item recognition	Item Classification	Average
PR2	Correlation	0.916	0.934	0.897	0.872	0.962	0.916
	MAE	0.024	0.026	0.037	0.024	0.025	0.027
	RMSE	0.033	0.036	0.054	0.035	0.034	0.038
GLM	Correlation	0.672	0.595	0.509	0.407	0.571	0.551
	MAE	0.050	0.064	0.088	0.053	0.081	0.067
	RMSE	0.061	0.081	0.105	0.065	0.101	0.083

Table 1. The Comparison of Prediction Models

These results give strong indication that such a prediction should be possible. Whilst there is an extremely high correlation between predicted and real performance in the PR2 model (average of 0.916), we suggest basing the conclusions on the GLM as it conforms to all underlying assumptions. The results show an average correlation of 0.551 with relatively small mean absolute and root mean squared errors (0.067 and 0.083). Note however that this preliminary analysis could be reporting over-fitted results as the numbers

above report the training set performance indicators rather than test set indicators gained, for example, via cross validation. Given the limited sample size we intend to use a leave-one-out-based cross validation method for our final analysis as well as employ other prediction methods such as regression trees.

H3a will be tested by comparing the error rates of the collaborative filtering and data imputation methods (H1b) with the error rates obtained from regression models (H2c). To establish statistical significance we will employ paired difference tests as Welch's t and Wilcoxon signed-rank test. When comparing the results of Table 1 and H1b above, we find that models based on cognitive capabilities (using the GLM) dominate over the collaborative filtering (employing SVDs): MAE_{GLM}=0.067 vs. MAE_{SVD}= 0.102 and $RMSE_{GLM}$=0.083 vs. $RMSE_{SVD}$=0.127, indicating that our hypothesis promises to be confirmed.

The evaluation of ***H3b*** will be most involved. The data collected per subject can be divided into the matrix *CO*, which contains all subjects' average result per ETS test in each cell, and the Matrix *CR*, which contains all subjects' average performance per crowd-task in each cell. For each subject *s* those two matrices, hence, contain a row with the respective results for this subject. This allows the computation of a generalized linear model to compute the transformation matrix *T*, to approximate the equation $CO * T = CR$. Note that $CO * T$ essentially specify a decomposition of *CR* along the cognitive dimensions specified by the chosen ETS tasks (the results of which are in the column's of *CO*). *T* now represents the mapping mentioned in ***H3b*,** which decomposes the performance on crowd-tasks to the cognitive capabilities.

To test ***H3b***, we will first learn *T* with the *CO* and *CR* for all but one subject *s*. We will then try to predict each crowdtasks' performance rating (which we will call missing the subsequent discussion) by transforming the other performance ratings (called known) of subject *s* by the means of *T* into its predicted cognitive capabilities. In other words, we will use *T* to decompose the subject's known crowd-task performances into its cognitive capabilities. Given that we now have the cognitive capabilities we can then use the regression contained in *T* to predict the value for the missing crowd-task performance. We then repeat this step for each all other crowd-tasks and subjects. We are well aware that this "round-trip" usage of *T* may increase noise. Note, however, that this usage is not unlike the usage of the decomposition matrices in singular value decomposition. The main difference is that the decomposition here is not based on mathematical properties of the matrix *CR*, but on the theoretical foundation of the cognitive dimensions defined by *CO*.

Limitations & Future Research

Our approach has the following noteworthy limitations. First, since we do not know the real distribution of the crowdworkers' cognitive capabilities the recruitment was done based on workers' ratings in a way that represented as much as possible a wide range of ratings. Nonetheless, the results seem to be lumped in a limited range of the performance scale. Given the limited sample size (n=37) it is unclear if this generalizes to whole population and extrapolate the prediction model to others or if the performances tend to shrink into observed limited range. Second, the experiment was designed in English, as most of the crowd-workers are English speakers. However, it misses the potential cultural differences among international crowd-workers. Third, as the subjects of the experiment are freelancers, it is likely that their performance is positively biased compared to crowd-workers in platforms such as Mechanical Turk, as they tend to be more professional and their motivation is not purely financial but also rely on reputation maintaining need. Fourth, given time constraints we did not include the whole set of cognitive tests but just a partial subset. In addition, the included cognitive tests were abbreviated to the first part. This hampers the generalization of the results to general ETS results. Finally, the absence of additional information about the subjects' backgrounds such as professional skills or English proficiency did not allow us to control for these variables.

In the future, we intend to complement our crowdsourcing platform experiment with a controlled laboratory experiment, where we can control for a wide range of confounding variables and can gather background information on the subjects. Second, we intend to extend our experiment from the freelancer-dominated Elance platform to other less professional work ethic dominated crowdsourcing markets such as Mechanical Turk. Last but not least, our experiments so far focused on visual tasks – one common type of crowd-tasks. In the future, we intend to explore other typical crowdsourced tasks such as translation, rewriting, transcription, content summarizing, or data entry.

Acknowledgments

This work was supported by the Swiss National Science Foundation under contract number 143411.

References

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), pp. 327-355.

Bandura, A. 2001. "Social Cognitive Theory: An Agentic Perspective," *Annual review of psychology* (52:1), pp. 1-26.

Bernstein, A., Klein, M., and Malone, T. W. 2012. "Programming the Global Brain," *Communications of the ACM* (55:5), pp. 41-43.

Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. 2010. "Soylent: A Word Processor with a Crowd Inside," *Proceedings of the 23nd annual ACM symposium on User interface software and technology*: ACM, pp. 313-322.

Breslow, N. 1996. "Generalized Linear Models: Checking Assumptions and Strengthening Conclusions," *Statistica Applicata* (8), pp. 23-41.

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), pp. 193-226.

Chiu, C.-M., Hsu, M.-H., 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), pp. 1872-1888.

Doan, A., Ramakrishnan, R., and Halevy, A. Y. 2011. "Crowdsourcing Systems on the World-Wide Web," *Communications of the ACM* (54:4), pp. 86-96.

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), pp. 195-209.

Dunnette, M. D. 1976. "Aptitudes, Abilities, and Skills," *Handbook of industrial and organizational psychology*), pp. 473-520.

Ekstrom, R. B., French, J. W., Harman, H. H., and Dermen, D. 1976. "Manual for Kit of Factor-Referenced Cognitive Tests," *Princeton, NJ: Educational Testing Service*).

Goodchild, M. F., and Glennon, J. A. 2010. "Crowdsourcing Geographic Information for Disaster Response: A Research Frontier," *International Journal of Digital Earth* (3:3), pp. 231-241.

Greengard, S. 2011. "Following the Crowd," *Communications of the ACM* (54:2), pp. 20-22.

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), pp. 389-399.

Horowitz, D., and Kamvar, S. D. 2010. "The Anatomy of a Large-Scale Social Search Engine," *Proceedings of the 19th international conference on World wide web*: ACM, pp. 431-440.

Hunter, J. E. 1986. "Cognitive Ability, Cognitive Aptitudes, Job Knowledge, and Job Performance," *Journal of vocational behavior* (29:3), pp. 340-362.

Jung, H. J. 2014. "Quality Assurance in Crowdsourcing Via Matrix Factorization Based Task Routing," *Proceedings of the companion publication of the 23rd international conference on World wide web companion*: International World Wide Web Conferences Steering Committee, pp. 3-8.

Kanfer, R., and Ackerman, P. L. 1989. "Motivation and Cognitive Abilities: An Integrative/Aptitude-Treatment Interaction Approach to Skill Acquisition," *Journal of applied psychology* (74:4), p. 657.

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., and Horton, J. 2013. "The Future of Crowd Work," *Proceedings of the 2013 conference on Computer supported cooperative work*: ACM, pp. 1301-1318.

Kristof, A. L. 1996. "Person-organization fit: An integrative review of its conceptualizations, measurement, and implications." *Personnel psychology* (49:1), pp. 1-49.

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), pp. 281-342.

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), pp. 151-163.

Mao, A., Kamar, E., and Horvitz, E. 2013. "Why Stop Now? Predicting Worker Engagement in Online Crowdsourcing," *First AAAI Conference on Human Computation and Crowdsourcing*.

Mason, W., and Watts, D. J. 2010. "Financial Incentives and the Performance of Crowds," *ACM SigKDD Explorations Newsletter* (11:2), pp. 100-108.

Mason, W. A., and Suri, S. "How to Use Mechanical Turk for Cognitive Science Research,").

Mayer, R. E. 2005. "Cognitive Theory of Multimedia Learning," *The Cambridge handbook of multimedia learning*), pp. 31-48.

Minder, P., Seuken, S., Bernstein, A., and Zollinger, M. 2012. "Crowdmanager-Combinatorial Allocation and Pricing of Crowdsourcing Tasks with Time Constraints," *Workshop on Social Computing and User Generated Content*, pp. 1-18.

Morris, R. R., Dontcheva, M., Finkelstein, A., and Gerber, E. 2013. "Affect and Creative Performance on Crowdsourcing Platforms," *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*: IEEE, pp. 67-72.

Morris, R. R., Dontcheva, M., and Gerber, E. M. 2012. "Priming for Better Performance in Microtask Crowdsourcing Environments," *Internet Computing, IEEE* (16:5), pp. 13-19.

Oleson, D., Sorokin, A., Laughlin, G. P., Hester, V., Le, J., and Biewald, L. 2011. "Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing," *Human computation* (11), p. 11.

Peters, T. A. 1996. "Human Factors in Information Systems: Emerging Theoretical Bases." Wiley Online Library.

Pinker, S. 1984. "Visual cognition: An introduction," *Cognition* (18:1), pp. 1-63.

Reinecke, K., and Bernstein A. 2013. "Knowing What a User Likes: A Design Science Ap-proach to Interfaces that Automatically Adapt to Culture," *MIS Quarterly* (37:2), pp. 427-453.

Rodrigues, A. 2014. "The Theory of Cognitive Dissonance: A Current Perspective," *Arquivos Brasileiros de Psicologia Aplicada* (22:2), pp. 126-127.

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), pp. 283-295.

Sculley, D., Malkin, R. G., Basu, S., and Bayardo, R. J. 2009. "Predicting Bounce Rates in Sponsored Search Advertisements," *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*: ACM, pp. 1325-1334.

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), pp. 337-360.

Sweller, J., Ayres, P., and Kalyuga, S. 2011. *Cognitive Load Theory*. Springer.

Tan, F. B., and Hunter, M. G. 2002. "The Repertory Grid Technique: A Method for the Study of Cognition in Information Systems," *MIS Quarterly* (26:1).

Vakharia, D., and Lease, M. 2013. "Beyond Amt: An Analysis of Crowd Work Platforms," *arXiv preprint arXiv:1310.1672*).

Velez, M. C., Silver, D., and Tremaine, M. 2005. "Understanding Visualization through Spatial Ability Differences," *Visualization, 2005. VIS 05. IEEE*: IEEE, pp. 511-518.

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), pp. 473-489.

Weiss, D. J., and Shanteau, J. 2003. "Empirical Assessment of Expertise," *Human Factors: The Journal of the Human Factors and Ergonomics Society* (45:1), pp. 104-116.

Weiss, D. J., and Shanteau, J. 2014. "Who's the Best? A Relativistic View of Expertise," *Applied Cognitive Psychology*).

White, I. R., Royston, P., and Wood, A. M. 2011. "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice," *Statistics in medicine* (30:4), pp. 377-399.

Whitla, P. 2009. "Crowdsourcing and Its Application in Marketing Activities," *Contemporary Management Research* (5:1).

Wilson, R. S., De Leon, C. F. M., Barnes, L. 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), pp. 742-748.

Zhang, H., Horvitz, E., Chen, Y., and Parkes, D. C. 2012. "Task Routing for Prediction Tasks," *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*: International Foundation for Autonomous Agents and Multiagent Systems, pp. 889-896.

Zhang, J., Ackerman, M. S., and Adamic, L. 2007. "Expertise Networks in Online Communities: Structure and Algorithms," *Proceedings of the 16th international conference on World Wide Web*: ACM, pp. 221-230.

Zhang, S., Wang, W., Ford, J., Makedon, F., and Pearlman, J. 2005. "Using Singular Value Decomposition Approximation for Collaborative Filtering," *E-Commerce Technology, 2005. CEC 2005. Seventh IEEE International Conference on*: IEEE, pp. 257-264.