Keep it Simple:
Reward and Task Design in Crowdsourcing

Ailbhe Finnerty
FBK-IRST
Via Sommarive, 18
38123, Povo, Italy.
finnerty@fbk.eu

Pavel Kucherbaev
DISI-UniTN
Via Sommarive, 5
38123, Povo, Italy.
kucherbaev@disi.unitn.it

Stefano Tranquillini
DISI-UniTN
Via Sommarive, 5
38123, Povo, Italy.
tranquillini@disi.unitn.it

Gregorio Convertino
Xerox Research Centre Europe
6 chemin de Maupertuis
38240 Meylan, Grenoble, France
gconvertino@gmail.com

ABSTRACT

Crowdsourcing is emerging as an effective method for performing tasks that require human abilities, such as tagging photos, transcribing handwriting and categorising data. Crowd workers perform small chunks of larger tasks in return for a reward, which is generally monetary. Reward can be one factor for motivating workers to produce higher quality results. Yet, as highlighted by previous research, the task design, in terms of its instructions and user interface, can also affect the workers' perception of the task, thus affecting the quality of results. In this study we investigate both factors, reward and task design, to better understand their role in relation to the quality of work in crowdsourcing. In Experiment 1 we test a variety of reward schemas while in Experiment 2 we measure the effects of the complexity of tasks and interface on attention. The long-term goal is to establish guidelines for designing tasks with the aim to maximize workers' performance.

AUTHOR KEYWORDS

Crowdsourcing, motivation, user interface, cognition

ACM CLASSIFICATION KEYWORDS

H.5.3. Group and Organization Interfaces: Web-based interaction.

INTRODUCTION

Crowdsourcing markets such as Amazon Mechanical Turk (AMT) are well known for providing fast results; however, they are not yet optimized for providing high-quality results [6]. Therefore, *understanding how to motivate workers to complete tasks with a high level of quality is extremely important for improving current crowdsourcing platforms* [8]. Motivation for improving quality can be extrinsic, giving a worker an explicit reward such as money, or intrinsic, such as a personal interest in the task itself, (e.g., a fun task, volunteering in an open source project).

Studies of worker performance on crowdsourcing platforms have shown how an economic reward leads to a higher output of results, yet, not of a high quality [11]. Mason and Watts have shown that increasing monetary rewards leads to faster, but not always better outputs [11].

In [12] the authors study how extrinsic and intrinsic motivations can improve output quality, they found that when no extrinsic motivation was given (no payment), intrinsic incentive did not play any role in motivating people. In [13] different incentives were used to improve the quality of the results, such as workers being asked to evaluate their peers' work, which led to better results, while financial incentives did not affect the quality. At the same time, in [3], it was shown that quality assessment feedback is well received by workers and can be useful for achieving better results.

The quality of the work is also related to the design of the task, not only the quality of the workers [9]. The design of tasks contains many aspects, from incentives to the interface and description. Task descriptions, which are clear to the requester, can be difficult for workers to understand and interfaces that are very complex for users can affect the task result quality [9]. Improving and changing the task design in terms of ergonomics and instructions can lead to better quality results.

EXPERIMENT 1

This experiment investigates different reward schemas, intrinsic and extrinsic motivations, in an attempt to improve both *time taken* to complete the task and *accuracy*.

Methodology

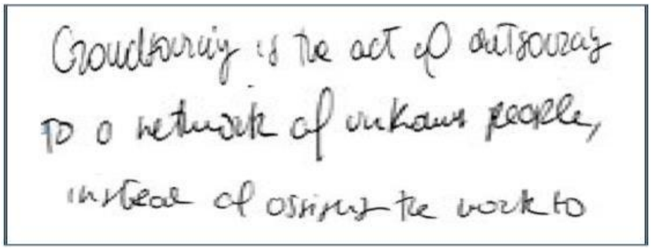
The task for the study was to correctly identify handwritten text and convert it to typed text. The sentence, a quote from [5], was "*crowdsourcing is the act of outsourcing the execution of work to a network of unknown people, instead of assigning the work to employees*" (Figure 1). The words were jumbled and the handwriting was slightly distorted to make the task more complex, the description of the task was always to "recognise all the words given on the pictures". The conditions for the study varied the reward schema for improving the time and the accuracy. The different rewards for the experiment were 1) *none*- when no extra reward was given for the task, 2) *please*- workers were asked to "please" do this task quicker or more accurately or 3) *fixed*- a reward of fixed amount which was given regardless of their performance and 4) *dynamic*- the reward was calculated based on their performance (in terms

CHItaly '13, September 16 - 20 2013, Trento, Italy

ACM 978-1-4503-2061-0/13/09…$15.00.
http://dx.doi.org/10.1145/2499149.2499168

of accuracy and time taken to complete the task), the better the performance the greater the reward.

This experiment used a full factorial design with two independent variables having 4 levels each (4x4 conditions). The two variables indicated the type of reward provided with respect to 1) *Time* (none, please, fixed bonus, dynamic bonus) and 2) *Accuracy* (none, please, fixed bonus, dynamic bonus), respectively. The experiment was run at two different times of the day to account for worker variability. We hypothesised that: **H1)** when the workers were motivated to produce quicker results, the time would improve from the baseline; **H2)** when the workers were motivated to provide better accuracy, it would improve above baseline; **H3)** when the workers were given multiple motivations or rewards (i.e., for time and accuracy), one motivation would take precedence and the motivation with the bigger reward would lead to a better performance.



Crowdsourcing is the act of outsourcing
to a network of unknown people,
instead of assigning the work to

Figure 1. Example of Handwriting.

Results

Using criteria established prior to the commencement of the experiments, due to large values of outliers, results one standard deviation above or below the mean, 23.5% of outliers were removed from the dataset (N=612). Since we had two blocks of runs, Time (1 and 2) was included as a covariate variable to control for confounding variability due to idiosyncratic samples.

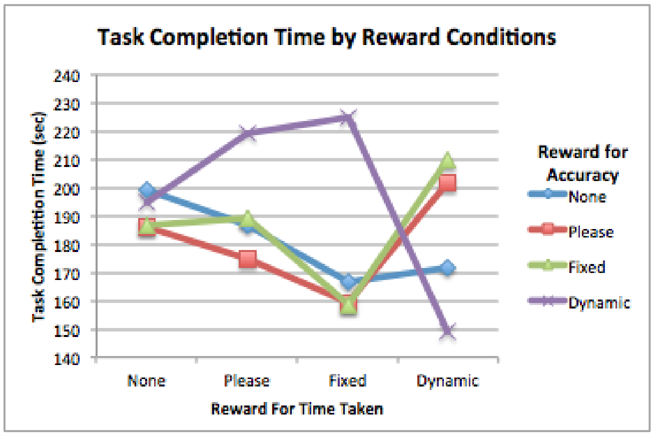


Figure 2. Interaction of Time Taken*Accuracy Motivation on workers response time (sec).

A Multivariate Analysis of Variance (MANOVA) was carried out on the data with two independent variables for the reward provided with respect to: 1) *Time* and 2) *Accuracy*; and two dependent variables: 1) *Time* (in seconds) and *Accuracy* (a percentage of correct answers). The analysis revealed no significant main effects of Time (M= 186.32, SD= 109.5) or Accuracy (M= 58.97, SD= 15.3). It did, however, reveal a significant interaction between the two manipulated variables (time and accuracy) on the dependent variable of time, F (9, 595) = 2.13, p <.05, but not on accuracy. The same interaction effect becomes even more evident when we consider only the conditions where both rewards are active (F (9,595) = 2.3, p = 0.01, (N=340)).

We found evidence of an interaction among the two co-existing reward schemes on the workers performance through both MANOVAs. The interaction is also visible in Figure 2 (the rightmost bullet of the purple line, which indicates dynamic bonus for both time and accuracy, provided the lowest average time). Moreover, when giving an extrinsic motivation or bonus-based rewards, we found that the hetero-scheme conditions (fixed-dynamic, dynamic-fixed) generally led to worse performances (higher average time taken) than homo-scheme conditions (dynamic-dynamic, fixed-fixed).

DISCUSSION

Dynamic-dynamic conditions led to the best performance in terms of time and accuracy. The results suggest that, as part of the task instructions, more complex descriptions of reward schemes for the worker to read and understand (i.e., what aspect of performance to optimize, time or accuracy) can lead to a reduced overall performance. The results did not support our initial hypothesis – H3. Based on our analyses of the differences across conditions, our interpretation is that an increased cognitive demand on the workers in attending to multiple motivational schemas at the same time is a key factor that explains the unexpected results. We plan to validate this new hypothesis in the next experiment by implementing a dual task design to manipulate the workers attention and examine the effects on their performance.

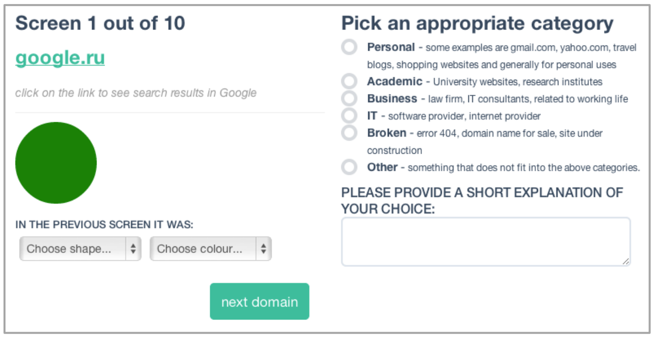
EXPERIMENT 2

The results of Experiment 1 suggest that when a worker is asked to provide faster and more accurate results in the same task, it is not clear to the worker which task parameter they should attend to and optimize (i.e., time or accuracy). According to Cognitive Load Theory [14], cognitive task-analysis methods [1] can be used to inform the design of experiments on crowdsourcing platforms to improve the quality of the results. Kittur, in [8] shows that it is possible to crowdsource complex tasks efficiently, if the task and tools given to the workers are designed properly. What seems critical is to design the task instructions and incentives so that are easy to understand and simple to follow.

This motivated us to investigate the effects of increased cognitive demand on the workers (i.e., divided attention), which correspond to the dual-reward schema in Experiment 1. In Experiment 2 we investigated complexity and cognitive demand on workers, by manipulating the task (*simple* vs. *dual task*) and the user interface (*simple* vs. *complex interface*). The task of the workers was to categorise Internet domain names (primary task) and at the same time they had to perform a (secondary) cognitive task. We expected that the performance of the primary task would suffer from the increased demands on the worker's attention. Orthogonally to this task manipulation, we also altered the complexity of the task interface. This second experiment helped us to better understand the findings of Experiment 1 and whether it was worth revisiting the idea of improving task performance by offering a dynamic reward as motivation for the workers.

Methodology

The task for Experiment 2 was to categorise Internet domain names (or URLs) into six possible categories: *1) Personal, 2) Academic, 3) Business, 4) IT, 5) Broken and 6) Other.* A short description of each category was given to guide the task. The full domain-categorisation task consisted of classifying ten domain names, for which workers were paid $0.15. For quality control, to ensure that the worker was not simply choosing random categories, an explanation of the rationale for the chosen category was given in a small text box.



Screen 1 out of 10

google.ru

click on the link to see search results in Google

IN THE PREVIOUS SCREEN IT WAS:

Choose shape... Choose colour...

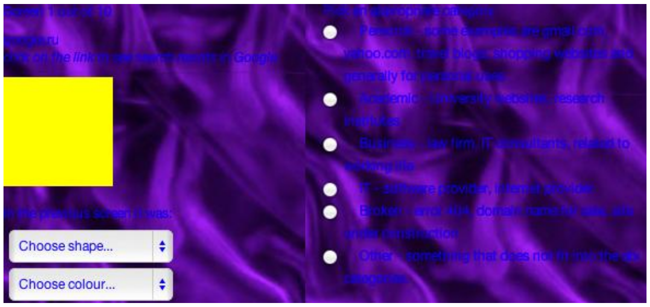
Pick an appropriate category

○ **Personal** - some examples are gmail.com, yahoo.com, travel blogs, shopping websites and generally for personal uses
○ **Academic** - University websites, research institutes
○ **Business** - law firm, IT consultants, related to working life
○ **IT** - software provider, internet provider
○ **Broken** - error 404, domain name for sale, site under construction
○ **Other** - something that does not fit into the above categories.

PLEASE PROVIDE A SHORT EXPLANATION OF YOUR CHOICE:

next domain

Figure 3. Simple User Interface



Screen 1 out of 10

google.ru

click on the link to see search results in Google

In the previous screen it was:

Choose shape...

Choose colour...

Pick an appropriate category

● Personal - some examples are gmail.com, yahoo.com, travel blogs, shopping websites and generally for personal uses
● Academic - University websites, research institutes
● Business - law firm, IT consultants, related to working life
● IT - software provider, internet provider
● Broken - error 404, domain name for sale, site under construction
● Other - something that does not fit into the above categories

Figure 4. Complex User Interface

For the dual task condition, performed in parallel with the domain-categorisation task, the workers were asked to remember the colour and the shape of a figure presented on the screen. On the first screen the worker was presented with the instructions and a coloured shape. They were instructed to remember the shape and colour as they would be asked about it in the next part of the task. By performing this second task the workers had to divide their attention while performing the domain-categorisation task: i.e., they had to simultaneously make a conscious effort to remember the shape and colour of the figure from the previous screen.

As well as investigating the effect of an increased cognitive load on workers, the effect of simple and complex user interfaces was investigated. For this condition we ran the task on two different User Interfaces (UI), simple and complex. For the simple UI we designed it with a clear layout and instructions (Figure 3) and for the complex UI (Figure 4) with a confusing background and unstructured layout. It has been found previously that complexity can contribute to a workers ability to perform a task well [7] and a badly designed UI can lead to a more complex task overall.

The study had a 2x2 design creating four conditions: 1.1 simple UI with dual task, 1.2 simple UI with no dual task, 2.1 complex UI with dual task and 2.2 complex UI with no dual task. We hypothesised that **H1)** the presence of a dual task, would put extra demand on the attention of the worker and lead to a worse overall performance than in the task without and **H2)** the complex UI would give worse results than the simple UI. If both hypotheses were verified then the conditions "1.2 simple UI with no dual task" and "2.1 complex UI with dual task" would lead to best and worst performance, respectively. We measured accuracy as agreement between workers: i.e., we assigned workers to each task in pairs and measured accuracy as percentage of agreement between the two workers in each pair.

Results

The statistical analysis found evidence in support of both of our hypotheses. A MANOVA was carried out on the data (80 pairs of participants, each sharing the same HIT, 160 runs in total) with two independent variables 1) *Task* (dual task, none) and 2) *User Interface* (simple, complex) and dependent variables 1) *Accuracy* (measured by the agreement between the workers), and 2) *Time taken to complete* the task (measured in seconds). Repeat Run was added as a covariate to the analysis to control for the learning effects of the workers who completed more than one task across the four conditions.

The MANOVA analysis revealed a main effect of task (F (2,153) = 6.39, p < .05) with better accuracy in the condition without dual task (M= 31.33 SD=14.17) than with dual task (M= 27.83, SD= 17.50). The main effect is also evidenced in the time taken to complete the task with lower completion time in the condition without dual task (M= 584.05s, SD= 350.4s) than with (M= 714.23s, SD= 564.3s). The analysis revealed an even more evident main effect of User Interface (F(2,153) = 3.97, p < .05) with better accuracy in the simple UI condition (M= 34.08, SD= 13.33) than the complex UI condition (M= 25.08, SD= 16.67) see Figure 5. Contrary to the finding for the User Interface main effect the completion time was actually quicker for the complex (M= 636.52s, SD= 472.9s) than for the simple condition (M= 660.6s, SD= 524.8s). While both effects were significant, the effect due to the manipulation on the UI complexity was stronger than the effect of the manipulation on the task complexity. The interaction between the two factors was not significant. The variability due to Repeat Run, or learning effect among repeated runs by the same participant, had been controlled for in the MANOVA analysis.

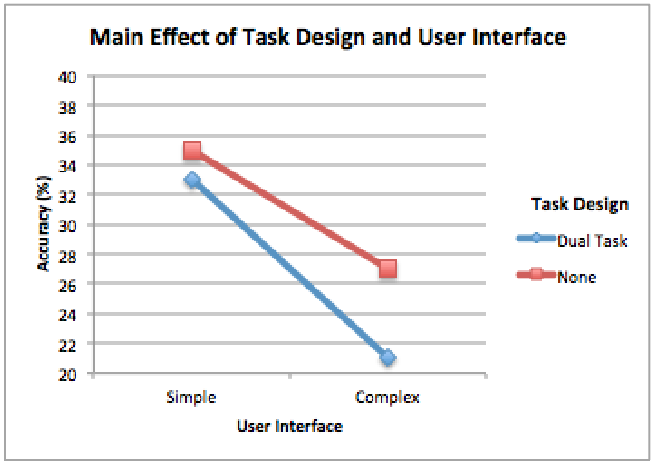


Figure 5. Main Effect of Task and User Interface.

Discussion

The results support our first hypothesis that the accuracy would be worse overall when having to perform a dual task (M's= 27.83 vs. 31.33). This supports our argument that having to attend to two separate commands at the same time can have the effect of increasing the cognitive demands on the worker, thereby having a detrimental effect on the performance of their work. Hypothesis 2, is also supported as the accuracy for the complex User Interface was significantly worse than for the simple (M's= 25.08 vs.34.08), showing evidence that by having a simple UI, it can make the task clearer for the worker to understand, which leads to better accuracy of the task. The phenomenon of increased cognitive load captured in Experiment 1, was replicated in more detail in Experiment 2. Finally, a learning effect was also found for the workers repeating the tasks in multiple conditions, which were controlled for in the MANOVA. This has implications for how to run future crowdsourcing studies: repeated runs and amount of prior experience of the worker should be measured and accounted for.

CONCLUSION

The aim of this study was to create a method for designing crowdsourcing tasks in order to produce more accurate results, focusing on providing motivation for the workers with different reward strategies. From Experiment 1 we found that dynamic reward as a motivation leads to better results, if applied to time and accuracy at the same time. More precisely, homogenous conditions (same reward for both dimensions) generally produce higher quality results. In addition to this finding, we discovered that a higher cognitive demand (divided attention factor), as well as a more complex description generally leads to a worse performance. In Experiment 2 we investigated complexity and cognitive demand on workers, by manipulating the task (simple vs. dual task) and the user interface (simple vs. complex interface). The results show evidence that a clearer and simpler design and less demand on workers' attention provide more accurate results. The findings of both experiments show evidence that keeping tasks simple leads to a better overall performance and suggests the need for further work on the design and implementation of tasks in crowdsourcing.

FUTURE WORK

We will continue this research toward the identification of specific patterns of task design and reward schemas that lead to better performances. Since different reward schemas and task designs can affect the performance in different ways, depending on the task nature (routine, algorithmic and creative), future experiments will systematically measure how different reward, task design, and task instructions influence the workers' performance for different types of tasks. The aim is to establish guidelines for designing tasks to maximize workers' performance.

REFERENCES

[1] Chipman, S. F., Schraagen, J. M., & Shalin, V. L. Introduction to cognitive task analysis. In J. M. Schraagen et.al. (Eds.), Cognitive task analysis (pp. 3-23). Mahwah, NJ: Lawrence Erlbaum Associates (2000).

[2] Crump MJC, McDonnell JV, Gureckis TM Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE* 8 (3). (2013).

[3] Dow, S., Kulkarni, A. P., Bunge, B., Nguyen, T., Klemmer, S. R., and Hartmann, B. Shepherding the crowd: managing and providing feedback to crowd workers. In *CHI Extend. Abstracts* (2011), 1669–1674.

[4] Heer, J. and Bostock, M. Crowdsourcing AMT graphical perception: using mechanical turk to assess visualization design. In *CHI* (2010), 203-212.

[5] Howe, J. The rise of crowdsourcing Wired Magazine 14 (6). (2006)

[6] Kaufmann, N., and Schulze, T. Worker motivation in crowdsourcing and human computation. *Education* 17, (2011).

[7] Kieras, D. and Polson P.G., An approach to the formal analysis of user complexity. *International Journal of Man-Machine Studies*, 22 (1985), 365-394.

[8] Kittur, A. Crowdsourcing, Collaboration and Creativity. *XRDS*, 17 (2010), 22-26.

[9] Kittur, A. Nickerson, J V. Bernstein, Ml Gerber, E. Shaw, A. Zimmerman, J. Lease, M. and Horton, J. The future of crowd work. In *CSCW* (2013)., 1301-1318.

[10] Komarov, S. Reinecke, K. and Gajos K. Z. Crowdsourcing Performance Evaluations of User Interfaces. In *CHI* (2013), 207-216.

[11] Mason, W. A., and Watts, D. J. Financial incentives and the "performance of crowds". In *KDD Workshop on Human Computation* (2009), 77–85.

[12] Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., and Vukovic, M. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *ICWSM* (2011).

[13] Shaw, A. D., Horton, J. J., and Chen, D. L. Designing incentives for inexpert human raters. In *CSCW* (2011), 275–284.

[14] Sweller, J. Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, 12 (1988) 257-85.