On Variety, Complexity, and Engagement in Crowdsourced Disaster Response Tasks

Sofia Eleni Spatharioti
Northeastern University
spatharioti.s@husky.neu.edu

Seth Cooper
Northeastern University
scooper@ccs.neu.com

ABSTRACT

Crowdsourcing is used to enlist workers as a resource for a variety of applications, including disaster response. However, simple tasks such as image labeling often feel monotonous and lead to worker disengagement. This provides a challenge for designing successful crowdsourcing systems. Existing research in the design of work indicates that task variety is a key factor in worker motivation. Therefore, we asked Amazon Mechanical Turk workers to complete a series of disaster response related subtasks, consisting of either image labeling or locating photographed areas on a map. We varied the frequency at which workers encountered the different subtask types, and found that switching subtask type at different frequencies impacted measures of worker engagement. This indicates that a certain amount of variety in subtasks may engage crowdsourcing workers better than uniform subtask types.

Keywords

crowdsourcing, Amazon Mechanical Turk, variety, complexity, engagement

INTRODUCTION

Crowdsourcing, in both paid and volunteer contexts, is an increasingly popular approach to acquiring and applying human skills for a wide variety of applications (Rouse 2010; Geiger et al. 2011; Schenk and Guittard 2011). Crowdsourcing is emerging as a promising method to involve the general public in disaster response. This lends itself to a variety of perspectives, such as studying social and organizational aspects of successful crowdsourced disaster response projects (Liu 2014; Goodchild and Glennon 2010; Crowley 2013) and examining the multitude of technical aspects related to crowdsourced disaster response (Imran et al. 2014; Lue et al. 2014), including crowdsourced processing of aerial imagery (Ofli et al. 2016; Munro et al. 2013).

To this end, various efforts have been established, with the main goal of exploring ways of involving the crowd in disaster response. Of particular relevance to our work is Public Lab, a community dedicated to developing open source hardware and software tools for collecting and analyzing data from environmental concerns, including disasters. One such tool was MapMill, developed by Public Lab and Jeff Warren, which was used for crowdsourced analysis of aerial images taken from the Gulf Coast after the Deepwater Horizon oil spill. MapMill was subsequently adapted for analysis of aerial images from Hurricane Sandy, taken by the Civil Air Patrol, in 2012 in a collaboration between the Humanitarian OpenStreetMap Team and FEMA (Munro et al. 2013). Volunteered data from MapMill were used to create a heatmap of the affected area, with applications for directing resources to areas with the highest damage reports (Meier 2012).

While both Hurricane Sandy and the Deepwater Horizon oil spill were large-scale events which attracted many volunteers and resources, this is not always the case for incidents in disaster response. Munro et al. (2013) pointed out the difficulty in recruiting volunteers for events with lesser magnitude, impact, and/or public outreach and suggested paid crowdsourcing as a cost-effective alternative. However, engaging paid workers or volunteers is key. In volunteer settings, it is known that most participants are only engaged for a short time and much of the work is accomplished by a small portion of people (Sauermann and Franzoni 2015; Sturn et al. 2015). In paid settings,

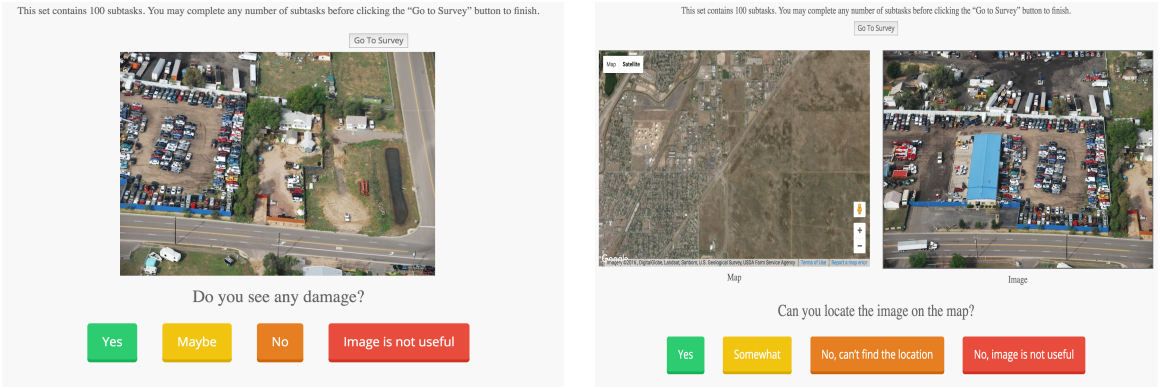

Figure 1. Example screenshots of the label (left) and map (right) subtasks.

worker motivations beyond simple payment are present (Kaufmann et al. 2011), and crowdsourcing marketplaces such as Amazon Mechanical Turk (MTurk) mean that another, more interesting task is only a click away.

As techniques for improving worker engagement and motivation in crowdsourced tasks are of interest, in this work, we explored the use of *variety* to improve crowdsourcing worker engagement in disaster response tasks. We focus primarily on *behavioral engagement*—that is, the observable persistence and effort one puts toward a task (Reeve 2015). We consider crowdsourcing tasks which can be broken down into a sequence of smaller subtasks, in which once one subtask is completed, the next subtask is served to the worker; this is a common workflow structure used in crowdsourcing. Thus, we use the term *subtask* to refer to one of these smaller, indivisible units of work, and *task* to refer to the overall sequence of subtasks served to a crowdsourcing worker.

We developed a task where workers were invited to complete a sequence of image related subtasks, specifically, *label subtasks*, a simple subtask where workers selected a label for an image from a set, and *map subtasks*, a more complex subtask where workers were asked to locate an image on a map and indicate if they were successful. When considering the *task complexity* (Wood 1986) of the subtasks, we view the map subtask as more complex as it contains, for example, both more "information cues" (the image and the map) and "required acts" (navigating the map and indicating success).

We posted a Human Intelligence Task (HIT) on MTurk that paid workers a fixed amount to complete any number of subtasks, with no minimum amount of work required, and answer a post-survey about their experience. We considered the baseline task to be a sequence of only label subtasks, similar to what a worker would encounter in MapMill. To examine the effects of variety—in terms of subtasks with different complexity—on performance, we varied the frequency with which the more complex map subtasks were served to workers during their progress through completing label subtasks. We wished to explore what impact, if any, differing schedules of task variety in our design would have on worker performance, including levels of engagement and output quality.

We found that when inserting map subtasks into a sequence of label subtasks at fixed intervals, certain intervals resulted in increased engagement—measured as voluntary time contributed to the task—compared to workers who only received label subtasks. In contrast, using only of one type of subtask—either maps or labels—did not lead to increased engagement when comparing one subtype to the other. Furthermore, introducing switches between the two different types of subtasks with related content did not observably affect workers' quality and speed on the label subtasks. We believe these findings indicate that tasks consisting of simple subtasks sequences can be made more engaging for workers by inserting more complex subtasks at appropriate intervals, without impacting quality. This work contributes to the growing literature on worker engagement in the design of crowdsourcing systems, by providing useful insights to worker performance in different settings of variety and complexity.

In the sections that follow, we start by presenting an overview of the related work that can be found in the present literature. We continue with providing a description of the task and experiment design. We then summarize our results analysis and conclude with some points of discussion.

RELATED WORK

In the past, there have been various efforts in involving volunteers in disaster response through crowdsourcing. Immediately after the Haiti earthquake of 2010, the Ushahidi project was established as a means of analyzing messages sent during the disaster and geolocating incidents in real-time (Liu 2014). Tomnod (Barrington et al.

2012) also invited volunteers to identify damage on buildings using post-earthquake satellite and aerial imagery of the Port-au-Prince area of Haiti. Volunteered geographic information has also been analyzed after the wildfires in Santa Barbara between 2007 and 2009 (Goodchild and Glennon 2010). Kerle et al. (Kerle and Hoffman 2013) addressed challenges that arise when designing a crowdsourcing task for disaster response, such as providing clear instructions that will ensure that volunteers will comprehend the task correctly and produce valid results, as well as generating useful maps for responders from the merged results. However, none of the above focused on exploring elements of task design to increase engagement.

Variety has long been recognized as a key factor in creating motivating work. Hackman and Oldham (1976) designed and tested a Job Characteristics Model, aiming to increase internal motivation and performance of employees through effective job design. They identified five core job characteristics, which include skill variety, along with task identity, task significance, autonomy, and job feedback. Lunenburg (2011) followed that work, providing an overview of empirical studies on the Job Characteristics Model, as well as applications of the model in the field of management. Other research has examined the tradeoffs between variety and specialization in work. Staats and Gino (2012) found that when examining the repetitive tasks of bank workers, specialization improved productivity over short periods of time (i.e., a day), while over longer periods, variety did. Narayanan et al. (2009) found that for software maintenance, a balance between specialization and variety led to the highest productivity.

In the crowdsourcing domain, Kittur et al. (2013), in their framework for future crowd work, proposed job design to support both organizational performance, but also worker satisfaction. Feedback from workers suggests that motivation is negatively impacted by monotonous tasks. Recent work has demonstrated that the ordering and continuity of crowdsourced microtasks can impact worker performance and engagement (Lasecki et al. 2014; Cai et al. 2016).

In work closely related to ours, Lasecki et al. (2015) examined the effects of contextual interruptions on crowdsourced microtasks (i.e. subtasks), where workers would switch between tasks of landmark locating on a map and image labeling. Their findings showed that switching context between closely related subtasks could slow workers down, but did not examine effects on workers' engagement. Dai et al. (2015) found that inserting short "micro-diversions" (such as pages of a graphic novel) into subtask sequences could improve worker engagement. However, these diversions did not result in work being completed. We see our work as a kind of combination of these two approaches: looking towards improving worker engagement through "diversions" that are different types of subtasks. Our work further explores the tradeoffs of various rates of interleaving different types of subtasks.

While we are primarily interested in worker engagement through variety, a wide body of approaches to improving crowdsourcing have been explored, including game mechanics (von Ahn and Dabbish 2004; Sturn et al. 2015), understanding payment schemes (Mason and Watts 2009; Gao and Parameswaran 2014; Acar and Ende 2011) and optimizing microtask workflows to reduce the amount of work needed to be done (Yan et al. 2011; Laws et al. 2011; Dai et al. 2010).

In this work, we used the total time workers spent voluntarily completing subtasks as an indication of engagement. When measuring engagement, a body of work related to ours has examined total time spent on task as a valid metric of engagement. In work by Khajah et al. (2016), MTurk was used to study "voluntary time on activity" as engagement. Participants were paid to try a game for several minutes, after which they could quit or continue to play voluntarily with no further compensation. Additional work has used time or "subtask" counts to operationalize engagement (Kassinove and Schare 2001; Andersen et al. 2011; D. Lomas et al. 2013; J. D. Lomas et al. 2016). In our work, workers were not required to spend any minimum time on the task or complete any minimum number of subtasks, as the setup allowed them to stop at any time, complete the survey questions, and get paid. Thus, we consider any time spent working on the task as voluntary and therefore a measure of engagement.

STUDY DESCRIPTION

Task Design

The subtasks we designed were inspired by the crowdsourced disaster response platform MapMill, where volunteers were asked to assess images from the Deepwater Horizon oil spill and Hurricane Sandy. Thus, our subtasks mainly involved viewing an aerial image and answering a question about the image by choosing from a small predefined set of answers. For this work we used a publicly available data set of Civil Air Patrol's aerial images of the 2013 Colorado floods, provided by the U.S. Geological Survey's Hazards Data Distribution System (2016), although workers were not informed of the source of the images. The two subtasks were the following:

- *Label subtasks*: Participants were presented with an image and were asked to answer a damage related question, with predefined answers. They were presented with a simple interface, containing an image, a question and buttons for four possible answers. The question chosen was *"Do you see any damage?"* and the

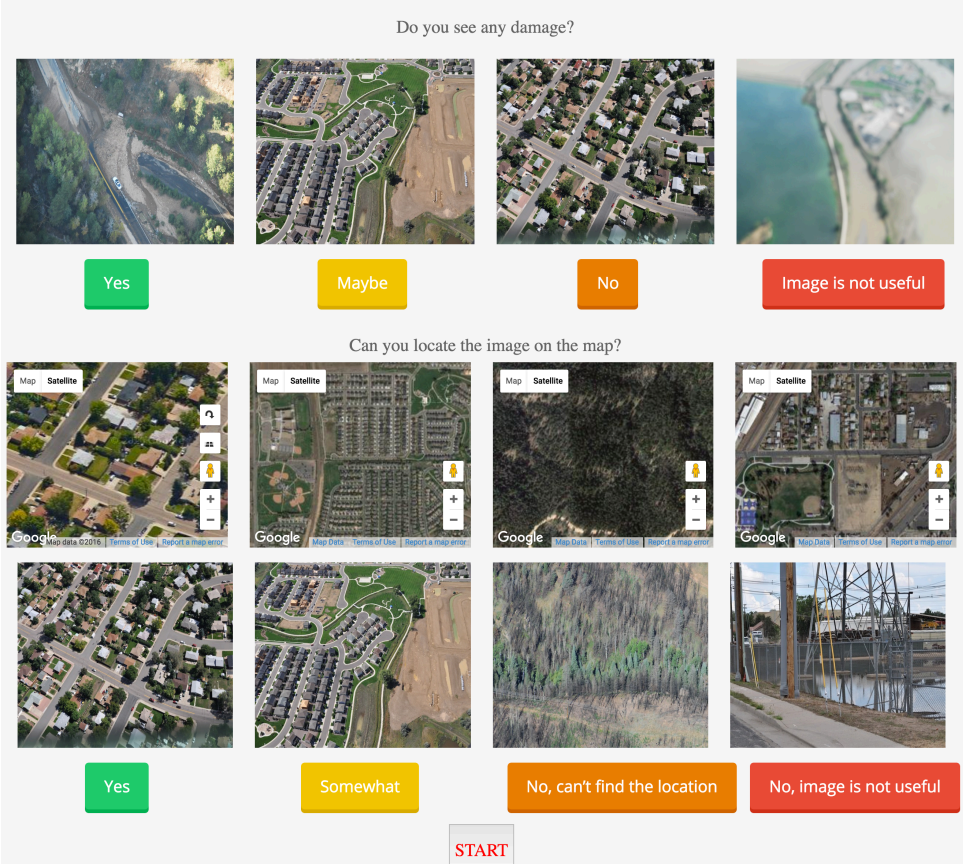

Figure 2. Example of the training provided to workers for our task.

possible answers were *"Yes"*, *"Maybe"*, *"No"* and *"Image is not useful"*. An example is shown in Figure 1 (left). We considered this to be a simple subtask; from a disaster response perspective, this type of subtask could be useful in locating areas of damage.

- *Map subtasks*: Alongside an image, participants had access to a map. provided by Google Maps (2016), and were asked to locate the image in the map view. Participants could pan and zoom on the map; the image's embedded GPS coordinates were used for the initial map location. The question chosen was *"Can you locate the image on the map?"* and the possible answers were *"Yes"*, *"Somewhat"*, *"No, can't find the location"* and *"No, Image is not useful"*. An example is shown in Figure 1 (right). We considered this to be a more complex subtask; from a disaster response perspective, this type of subtask could be useful for confirming and improving the locations of images.

Experiment Design

We ran a HIT on MTurk to recruit participants to complete subtasks. We chose MTurk for participant recruitment as it remains one of the most popular crowdsourcing marketplaces, is widely used in crowdsourcing research, and has a ready pool of participants who can be rapidly recruited. Additionally, MTurk workers have been shown to behave similarly to "traditional" subject pools (Paolacci et al. 2010) and to be motivated by more than payment (Mason and Watts 2009; Kaufmann et al. 2011; Chandler and Kapelner 2013). IRB approval was received for the purposes of the study. The HIT paid 50¢. The HIT title, which informed workers that they would be doing work related to disasters, was:

> Disaster Area Map Image Tasks

The HIT description was:

> Perform image related subtasks such as image labeling and image locating on a map and answer a short survey on the task.

The HIT keywords were:

> image, locate, locating, center, map, labeling, label

Upon accepting the HIT, workers were informed they were participating in a study and required to consent to proceed. Workers were then provided with the following general instructions:

> Click **START** to begin completing a series of image related tasks. For each image, answer the question that appears beneath the image. Some tasks require locating the images on the right on the map on the left. You may move around, zoom and, if available, rotate the map view, using the corresponding map controls. Try to center the map on the location shown in the image, making the map look as much like the image as possible. Please do not use your browser's back button while locating or taking the survey. This HIT requires Javascript. When you are finished, click the "Go To Survey" button to complete a short survey, after which your HIT will be completed. **This set contains 100 subtasks. You may complete any number of subtasks before clicking the "Go to Survey" button to finish. If you complete the survey, your submission will be approved.**

Motivated by the work of Ho et al. (2015), to mitigate effects of different beliefs workers may bring about the amount of work required to be approved, the instructions explicitly state submissions will be approved if the survey is completed.

After an example for the types of subtasks, which acted as a small tutorial (shown in Figure 2), workers then proceeded to complete subtasks by answering the questions that accompanied the images. In our HIT, workers were allowed to complete *up to* 100 subtasks.

At any point, participants could either continue completing subtasks or choose to complete the HIT by answering a short survey. If a worker completed all possible subtasks, they were automatically taken to the survey. In the survey, participants were asked to provide some feedback about the reasons that led them to their performance, in the form of a multiple selection question. The survey asked workers, *"Can you tell us why did you complete the number of subtasks that you did? Check all that apply."* Workers could check any number of the following checkboxes: *"I was engaged in the task," "I thought my submission would not get approved," "I wanted to help the project," "I did not understand the instructions," "I had no indication to stop," "I thought the payment was worth that much work," "I thought I would get paid more," "I wanted to see more images,"* and *"Other"*.

While there was a maximum of 100 subtasks available, workers were able to finish working on the subtasks at any time without penalty. Workers would be paid the fixed reward for completing the HIT, regardless of the number of subtasks completed—even if they completed no work at all—so we consider the work done and time spent on it essentially voluntary. Thus we believe examining work done and time spent are reasonable measures of *behavioral* engagement, as workers can, at any time, finish the HIT by going to the survey to collect their payment and move on to another HIT they feel is more worthwhile. As discussed in the related work, other work has used measures of time to examine engagement on MTurk.

We carried out a between subjects experiment design, recruiting 720 workers who completed the HIT, with an additional 133 workers who started, but did not complete, the HIT. Workers were randomly assigned into one of 6 conditions with different proportions of subtasks. In the `AllL` and `AllM` *uniform conditions*, workers were served all label or all map subtasks, respectively. In the `4L:1M`, `9L:1M`, `19L:1M`, and `29L:1M` *variety conditions*, workers were served one map subtask at a regular interval of label subtasks (e.g., in `4L:1M` one map subtask was served after every 4 label subtasks). As workers were randomly assigned into a condition, they did not have any control over the type of subtasks they were given. The images used in the subtasks were randomly ordered for each subject. Aside from the different proportions of subtask types, all workers received exactly the same HIT. For example, workers would receive the example map subtask tutorial screen even if they were not served any actual map subtasks. MTurk did not allow participants to participate in the study more than once, which ensures that our data does not contain duplicate entries for workers.

RESULTS

Data Analysis

We analyzed the following variables of worker performance on the subtasks:

- *Subtask Count*: The total number of subtasks, both label and map type, completed.
- *Label Count*: The total number of label type subtasks completed.
- *Subtask Agreement*: The percentage of all types of subtasks, both label and map, whose answers agree with the consensus answer. As we do not have ground truth for our subtasks, we use the consensus answer as the "correct" response. This was simply the answer selected by the majority of workers. We use agreement to evaluate worker accuracy.
- *Label Agreement*: The percentage of label type subtasks whose answers agree with the consensus answer, used as ground truth.

Variable	AllL ~ AllM	AllL ~ 4L:1M	AllL ~ 9L:1M	AllL ~ 19L:1M	AllL ~ 29L:1M
Subtask Count $p < .001$	**20 ~ 6** $p < .001, r = 0.55$	*20 ~ 15* $p = .052, r = 0.18$	20 ~ 22	20 ~ 17	20 ~ 20
Label Count $p < .001$	—	**20 ~ 12** $p < .001, r = 0.29$	20 ~ 19	20 ~ 16	20 ~ 19
Subtask Agreement $p = .002$	**75% ~ 67%** $p = .042, r = 0.20$	75% ~ 73%	75% ~ 76%	75% ~ 75%	75% ~ 80%
Label Agreement *n.s.*	—	73% ~ 75%	73% ~ 78%	73% ~ 76%	73% ~ 80%
Subtask Time $p < .001$	**8s ~ 26s** $p < .001, r = 0.70$	**8s ~ 16s** $p < .001, r = 0.51$	**8s ~ 13s** $p < .001, r = 0.39$	**8s ~ 11s** $p < .001, r = 0.28$	8s ~ 8s
Label Time $p = .05$	—	8s ~ 10s	8s ~ 9s	8s ~ 9s	8s ~ 7s
Total Time $p = .021$	193s ~ 179s	**193s ~ 343s** $p = .033, r = 0.19$	**193s ~ 285s** $p = .021, r = 0.21$	193s ~ 211s	193s ~ 205s
Abandonment $p = .014$	**8% ~ 22%** $p = .007, \phi = 0.03$	8% ~ 15%	*8% ~ 18%* $p = .072, \phi = 0.15$	*8% ~ 18%* $p = .082, \phi = 0.15$	8% ~ 12%
Understand $p = .036$	6% ~ 12%	6% ~ 7%	6% ~ 4%	6% ~ 2%	6% ~ 7%

Table 1. Summary of data and statistical comparisons. *p*-values given in the first column are for omnibus tests; in other columns, for post-hoc tests. Significant post-hoc comparisons highlighted in bold, borderline significant post-hoc comparisons highlighted in italics. Numerical variables (including *Subtask Agreement* and *Label Agreement*) are given as medians and Boolean variables as percentages.

- *Subtask Time*: The average time spent (in seconds) on each subtask for all subtasks completed, both label and map type.
- *Label Time*: The average time spent (in seconds) on each label type subtask. If workers are impacted by switching context between subtask types, they may become less efficient and take more time to complete label subtasks.
- *Total Time*: Total time spent on subtasks (in seconds). This variable was calculated as the time between seeing the first subtask and moving on to the survey.
- *Abandonment*: The percentage of workers who abandoned (accepted but did not complete) the HIT.

For time-related variables, any idle blocks of time 5 minutes or greater were removed from consideration. To identify significant differences, we used Pearson's chi-squared test for Boolean variables (*Abandonment* and survey responses) and Kruskal-Wallis test for numerical variables, as these were not normally distributed. If the omnibus test was significant, we then performed post-hoc pairwise comparisons between each condition and the `AllL` uniform condition, using Pearson's chi-squared test for Boolean variables and Wilcoxon rank-sum test for numerical variables. Pairwise tests were adjusted using the Bonferroni correction, scaling p-values by 5, when all possible comparisons were applicable, or by 4, for applicable comparisons. Unless otherwise noted, we report numerical averages as medians due to non-normality. To compute effect sizes, we used rank-biserial correlation (r) for numerical variables and phi (ϕ) for Boolean variables.

Given the potentially large number of post-hoc pairwise comparisons between conditions, we focused on two comparison groups:

- `AllL` against `AllM`. Comparing the two uniform conditions allows us to examine the impact of the different subtask types in isolation.
- `AllL` against each of the variety conditions containing both labels and maps; i.e., comparing a sequence of only label subtasks to inserting map subtasks into a label subtask sequence at different intervals.

Table 1 contains a summary of the variables and statistical comparisons. Figure 3 shows a survival chart, displaying worker retention over time for each of the conditions.

We found no significant differences in the survey responses other than for the response *"I did not understand the instructions"*, which we refer to as the variable *Understand*. Response rates for the other choices were: *"I was engaged in the task"* 36%, *"I thought my submission would not get approved"* 10%, *"I wanted to help the project"*

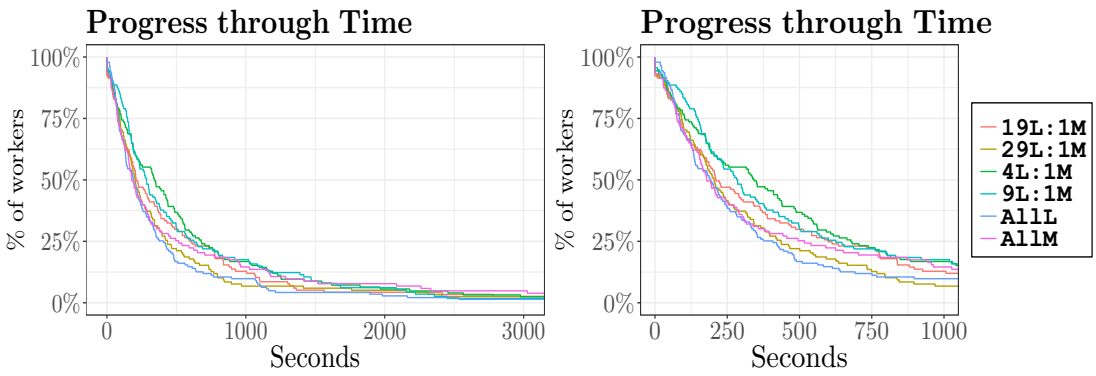

Figure 3. Survival chart for worker retention over time in the different conditions. The x-axis shows progress through time and the y-axis shows the percentage of workers retained up to that point; left shows up to 3000 seconds and right shows detail up to 1000 seconds. Note that early on, `4L:1M` and `9L:1M` retain more workers than conditions that serve map subtasks more or less frequently.

44%, *"I had no indication to stop"* 19%, *"I thought the payment was worth that much work"* 20%, *"I thought I would get paid more"* 8%, *"I wanted to see more images"* 28%, and *"Other"* 15%.

Findings

We summarize our findings in the following points:

The map subtask was generally more complex than the label subtask. Looking at the `AllL` against `AllM` comparison allows us to compare worker behavior on each of the subtask types themselves. The increased *Subtask Time* workers spent working on a single map subtask is consistent with the increased time expected with increased complexity (Campbell and Gingrich 1986). Additionally, among workers in the `AllM` condition, there was decreased *Subtask Agreement* and *Subtask Count*, along with increased in *Abandonment* of the HIT, compared to the `AllL` condition. Of note, though not statistically significant in all cases, is the observation that the `AllM` condition had the highest rate of abandonment and indication of reporting issues understanding the instructions. We do not consider this a surprising insight, rather supporting our initial assumption about the difference in complexity between subtask types.

Using only one type of subtask or the other did not impact total time spent on the task. Also looking at the `AllL` against `AllM` comparison, there was no significant difference in *Total Time* between the uniform subtask conditions, indicating that neither individual subtask type was inherently more engaging, with respect to time spent, than the other.

Interleaving subtask types did not observably impact speed or quality through distractions. When comparing `AllL` against each of the conditions with subtask variety, no significant pairwise comparisons were found among *Label Time* or *Label Agreement*. This indicates that workers remained focused on the task and the quality of their work was not negatively impacted by the introduction of subtask type switches.

In some conditions, inserting map subtasks into a label subtask sequence increased the time spent on the task. Again comparing `AllL` against the subtask variety conditions, workers in the `4L:1M` and `9L:1M` conditions spent significantly more *voluntary* time on the task, when compared to the `AllL` condition. However, the `4L:1M` condition was combined with a significant reduction in label count, while the `9L:1M` condition did not show a negative impact on label count. Also of note, though not statistically significant in all cases, is the observation that uniform subtask conditions (`AllL` and `AllM`) had the lowest *Total Time* among all conditions.

There was little, if any, observable impact on workers' subjective experience between conditions. There were no significant differences in the various survey responses among conditions. Even *Understand*, which was significant in the omnibus test, did not show significant differences in the post-hoc tests. This indicates that we did not observe a difference in workers' subjective experience and opinions regarding *self-reported* engagement, concern about getting approved, or the amount of work worth doing for the payment, and so forth. Despite this, we did observe differences in workers' actual behavior. However, it is of note that the top two selected survey responses indicated wanting to help the project, feeling engaged, and wanting to see more images, and the lowest were not understanding the instructions and desiring more pay, across all conditions. This provides some additional support to workers participating voluntarily even thought they were receiving payment.

We believe these observations indicate that the comparison between `AllL` and `9L:1M` is of particular interest. When compared with a sequence of simple subtasks (labels), interleaving a more complex subtask (maps) at a small interval increased the total time *voluntarily* spent on the HIT without observably negatively impacting other performance variables *of those workers who completed the HIT*. Figure 3 highlights the increased portion of workers in variety conditions with maps frequently interleaved who are retained early on. Examining other variety conditions indicates that the interval at which the map subtasks are interleaved matters. As the median *Subtask Count* was generally around 20 for conditions with many labels, it is possible that in the conditions with more space between the map subtasks, workers simply did not encounter enough maps to affect our measures, or most workers finished before even reaching a map.

DISCUSSION

In this study, we explored the effects of the variety of subtasks with differing complexity in crowdsourced disaster response, by inviting workers to complete a sequence of label and map subtasks—inspired by the MapMill disaster response project—served at various levels of frequency. We recruited paid workers through an MTurk HIT.

Our findings offer some interesting insights on how to design engaging tasks for disaster response. We believe that task design is important, not only for recruiting more volunteers, but also for improving retention and performance, in events that will otherwise face difficulty in attracting response participants. The increase in data volume can, in turn, facilitate more rapid deployment of response and recovery in areas struck by a disaster. This work constitutes a preliminary step into designing a platform for attracting participants to engage with disaster response efforts.

Although the results of our study provide initial support to the benefits of task variety in the design of crowdsourcing tasks, further study can help inform a deeper understanding of how different levels of variety and complexity affect worker performance and engagement.

Worker performance indicated a higher level of difficulty with the map subtask, which may have also contributed to lower rates of agreement and increased abandonment when map subtasks were frequent. Further work can explore lessening the rate of abandonment when workers are given such complex tasks.

This work examined only two specific types of subtasks; image labeling and locating images on map. Future work could be centered on expanding research to include other types of subtasks that are often found on crowdsourcing platforms like MTurk, such as manuscript transcription or sentiment analysis. It would also be interesting to explore variety among more than two types of subtask. We also only considered fixed intervals of providing the more complex map subtasks; in particular, variable or adaptive approaches to scheduling variety may be a fruitful area of exploration.

Finally, although we used voluntary time as a measure of behavioral engagement, we carried out our study in the context of paid crowdsourcing on MTurk, where participants were paid for participating. It remains to be confirmed that a similar effect would be present in purely voluntary crowdsourcing contexts.

ACKNOWLEDGMENTS

This work was supported by a Northeastern University TIER 1 grant, Google, and the New World Foundation in collaboration with Public Lab. We would like to thank the Civil Air Patrol and the U.S. Geological Survey for making the images available and the Mechanical Turk workers for their participation.

REFERENCES

Acar, O. A. and Ende, J. van den (2011). "Motivation, reward size and contribution in idea crowdsourcing". In: *DIME-DRUID ACADEMY*.

von Ahn, L. and Dabbish, L. (2004). "Labeling images with a computer game". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 319–326.

Andersen, E., Liu, Y.-E., Snider, R., Szeto, R., and Popović, Z. (2011). "Placing a Value on Aesthetics in Online Casual Games". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1275–1278.

Barrington, L., Ghosh, S., Greene, M., Har-Noy, S., Berger, J., Gill, S., Lin, A. Y.-M., and Huyck, C. (2012). "Crowdsourcing earthquake damage assessment using remote sensing imagery". In: *Annals of Geophysics* 54.6.

Cai, C. J., Iqbal, S. T., and Teevan, J. (2016). "Chain reactions: the impact of order on microtask chains". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3143–3154.

Campbell, D. J. and Gingrich, K. F. (1986). "The interactive effects of task complexity and participation on task performance: a field experiment". In: *Organizational Behavior and Human Decision Processes* 38.2, pp. 162–180.

Chandler, D. and Kapelner, A. (2013). "Breaking monotony with meaning: motivation in crowdsourcing markets". In: *Journal of Economic Behavior & Organization* 90, pp. 123–133.

Crowley, J. (2013). "Connecting grassroots and government for disaster response". In: *Commons Lab of the Woodrow Wilson International Center for Scholars*.

Dai, P., Mausam, and Weld, D. S. (2010). "Decision-theoretic control of crowd-sourced workflows". In: *Proceedings of the 24th AAAI Conference on Artificial Intelligence*.

Dai, P., Rzeszotarski, J. M., Paritosh, P., and Chi, E. H. (2015). "And now for something completely different: improving crowdsourcing workflows with micro-diversions". In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pp. 628–638.

Gao, Y. and Parameswaran, A. (2014). "Finish them!: pricing algorithms for human computation". In: *Proceedings of the VLDB Endowment* 7.14, pp. 1965–1976.

Geiger, D., Seedorf, S., Schulze, T., Nickerson, R. C., and Schader, M. (2011). "Managing the crowd: towards a taxonomy of crowdsourcing processes". In: *Proceedings of the Americas Conference on Information Systems*.

Goodchild, M. F. and Glennon, J. A. (2010). "Crowdsourcing geographic information for disaster response: a research frontier". In: *International Journal of Digital Earth* 3.3, pp. 231–241.

Google Maps (2016). http://maps.google.com/.

Hackman, J. R. and Oldham, G. R. (1976). "Motivation through the design of work: test of a theory". In: *Organizational Behavior and Human Performance* 16.2, pp. 250–279.

Hazards Data Distribution System Explorer (2016). http://hddsexplorer.usgs.gov/.

Ho, C.-J., Slivkins, A., Suri, S., and Vaughan, J. W. (2015). "Incentivizing high quality crowdwork". In: *Proceedings of the 24th International Conference on World Wide Web*, pp. 419–429.

Imran, M., Castillo, C., Lucas, J., Meier, P., and Vieweg, S. (2014). "AIDR: Artificial Intelligence for Disaster Response". In: *Proceedings of the 23rd International Conference on World Wide Web*, pp. 159–162.

Kassinove, J. I. and Schare, M. L. (2001). "Effects of the "near miss" and the "big win" on persistence at slot machine gambling". eng. In: *Psychology of Addictive Behaviors: Journal of the Society of Psychologists in Addictive Behaviors* 15.2, pp. 155–158.

Kaufmann, N., Schulze, T., and Veit, D. (2011). "More than fun and money. Worker motivation in crowdsourcing – a study on Mechanical Turk". In: *Proceedings of the Americas Conference on Information Systems*.

Kerle, D. N. and Hoffman, R. (2013). "Collaborative damage mapping for emergency response : the role of Cognitive Systems Engineering". In: *Natural hazards and earth system sciences* 13.1, pp. 97–113.

Khajah, M. M., Roads, B. D., Lindsey, R. V., Liu, Y.-E., and Mozer, M. C. (2016). "Designing engaging games using Bayesian optimization". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5571–5582.

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*, pp. 1301–1318.

Lasecki, W. S., Marcus, A., Rzeszotarski, J. M., and Bigham, J. P. (2014). *Using microtask continuity to improve crowdsourcing*. Tech. rep. CMU-HCII-14-100. School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Lasecki, W. S., Rzeszotarski, J. M., Marcus, A., and Bigham, J. P. (2015). "The effects of sequence and delay on crowd work". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1375–1378.

Laws, F., Scheible, C., and Schütze, H. (2011). "Active learning with Amazon Mechanical Turk". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1546–1556.

Liu, S. B. (2014). "Crisis crowdsourcing framework: designing strategic configurations of crowdsourcing for the emergency management domain". In: *Computer Supported Cooperative Work* 23.4-6, pp. 389–443.

Lomas, D., Patel, K., Forlizzi, J. L., and Koedinger, K. R. (2013). "Optimizing Challenge in an Educational Game Using Large-scale Design Experiments". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 89–98.

Lomas, J. D., Forlizzi, J., Poonwala, N., Patel, N., Shodhan, S., Patel, K., Koedinger, K., and Brunskill, E. (2016). "Interface Design Optimization As a Multi-Armed Bandit Problem". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4142–4153.

Lue, E., Wilson, J. P., and Curtis, A. (2014). "Conducting disaster damage assessments with Spatial Video, experts, and citizens". In: *Applied Geography* 52, pp. 46–54.

Lunenburg, F. C. (2011). "Motivating by enriching jobs to make them more interesting and challenging". In: *International Journal of Management, Business, and Administration* 15.1, pp. 1–11.

Mason, W. and Watts, D. J. (2009). "Financial incentives and the "performance of crowds"". In: *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pp. 77–85.

Meier, P. (2012). *Crowdsourcing the evaluation of post-Sandy building damage using aerial imagery*. https://irevolutions.org/2012/11/01/crowdsourcing-sandy-building-damage/.

Munro, R., Schnoebelen, T., and Erle, S. (2013). "Quality analysis after action report for the crowdsourced aerial imagery assessment following Hurricane Sandy". In: *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management*.

Narayanan, S., Balasubramanian, S., and Swaminathan, J. M. (2009). "A matter of balance: specialization, task variety, and individual learning in a software maintenance environment". In: *Management Science* 55.11, pp. 1861–1876.

Ofli, F., Meier, P., Imran, M., Castillo, C., Tuia, D., Rey, N., Briant, J., Millet, P., Reinhard, F., Parkan, M., et al. (2016). "Combining human computing and machine learning to make sense of big (aerial) data for disaster response". In: *Big Data* 4.1, pp. 47–59.

Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). "Running experiments on Amazon Mechanical Turk". In: *Judgment and Decision Making* 5.5, pp. 411–419.

Reeve, J. (2015). *Understanding Motivation and Emotion (Sixth edition)*. Hoboken, New Jersey: Wiley.

Rouse, A. C. (2010). "A preliminary taxonomy of crowdsourcing". In: *Proceedings of the 21st Australasian Conference on Information Systems*.

Sauermann, H. and Franzoni, C. (2015). "Crowd science user contribution patterns and their implications". In: *Proceedings of the National Academy of Sciences* 112.3, pp. 679–684.

Schenk, E. and Guittard, C. (2011). "Towards a characterization of crowdsourcing practices". In: *Journal of Innovation Economics & Management* 1, pp. 93–107.

Staats, B. R. and Gino, F. (2012). "Specialization and variety in repetitive tasks: evidence from a Japanese bank". In: *Management Science* 58.6, pp. 1141–1159.

Sturn, T., Wimmer, M., Salk, C., Perger, C., See, L., and Fritz, S. (2015). "Cropland Capture – a game for improving global cropland maps". In: *Proceedings of the 10th International Conference on the Foundations of Digital Games*.

Wood, R. E. (1986). "Task complexity: definition of the construct". In: *Organizational Behavior and Human Decision Processes* 37.1, pp. 60–82.

Yan, Y., Rosales, R., Fung, G. M., and Dy, J. G. (2011). "Active learning from crowds". In: *Proceedings of the 28th International Conference on Machine Learning*, pp. 1161–1168.