A Required Work Payment Scheme for Crowdsourced Disaster Response: Worker Performance and Motivations

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

Rebecca Govoni
Northeastern University
govoni.r@husky.neu.edu

Jennifer S. Carrera
Michigan State University
jcarrera@msu.edu

Sara Wylie
Northeastern University
s.wylie@northeastern.edu

Seth Cooper
Northeastern University
scooper@ccs.neu.com

ABSTRACT

Crowdsourcing is an increasingly popular approach for processing data in response to disasters. While volunteer crowdsourcing may suffice for high-profile disasters, paid crowdsourcing may be necessary to recruit workers for less prominent events. Thus, understanding the impact of payment schemes on worker behavior and motivation may improve outcomes. In this work, we presented workers recruited from Amazon Mechanical Turk with a disaster response task in which they could provide a variable number of image ratings. We paid workers a fixed amount to provide a minimum number of image ratings, allowing them to voluntarily provide more if desired; this allowed us to examine the impact of different amounts of *required work*. We found that requiring *no* ratings resulted in workers voluntary completing *more* work, and being more likely to indicate motivation related to interest on a post survey, than when small numbers of ratings were required. This is consistent with the motivational *crowding-out* effect, even in paid crowdsourcing. We additionally found that providing feedback on progress positively impacted the amount of work done.

Keywords

crowdsourcing, Amazon Mechanical Turk, payment, motivation, required work

INTRODUCTION

When faced with a large amount of data that would be either computationally challenging or rely on human subjectivity to process, crowdsourcing is a popular approach to gathering information on solutions provided by humans. In particular, image analysis tasks—such as image labeling or rating—have proven to be particularly amenable to this approach. A number of projects over the past decade have taken crowdsourced approaches to acquiring labels for images, either for the sake of acquiring the labels themselves or as training data for machine learning techniques (von Ahn and Dabbish 2004; Raddick et al. 2010; Mitry et al. 2013). Within this area, crowdsourced image rating for mapping—especially in the context of disaster response—has recently arisen. Projects such as SandyMill[1] (a collaboration between the Humanitarian OpenStreetMap Team and FEMA) (Munro et al. 2013), Tomnod's involvement in GEO-CAN (Barrington et al. 2012), the Ushahidi-Haiti Project (Liu 2014), and Cropland Capture (Sturn et al. 2015) have all taken approaches to asking crowd workers to rate images for

[1]SandyMill forked an existing open source image sorting application, MapMill, developed by Public Lab and Jeff Warren. In SandyMill, crowd members sorted images taken by Civil Air Patrol of damage from Hurricane Sandy.

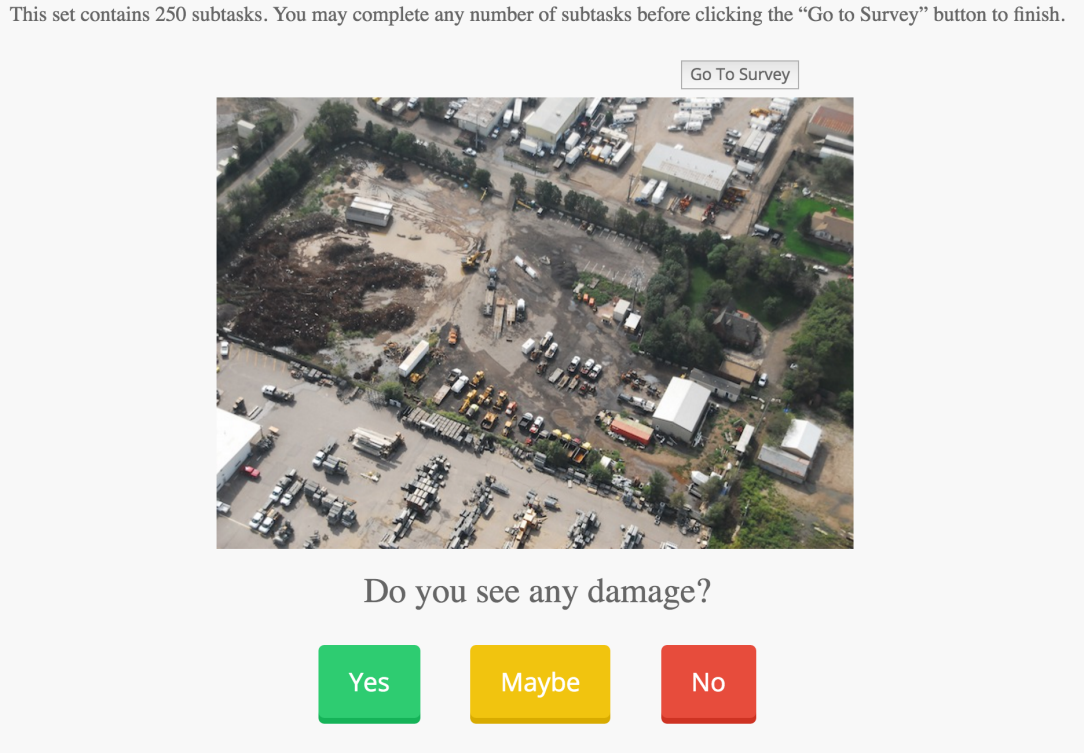

Figure 1. Screenshot of the image rating page used for our task. Instructions are shown along the top. In this condition, progress through the total set and progress through the required work are not available to workers.

the purposes of creating, improving, or annotating maps. "Space archaeology" is another emerging area for crowdsourcing image analysis, with projects such as Global Xplorer, a platform for analyzing satellite images (Global Xplorer 2016).

Crowdsourcing holds promise for applying human processing to the massive amounts of data that can be generated in the wake of disasters—either through paid work or volunteering. However, analysis of contribution patterns of participants in volunteer crowdsourcing projects generally indicate that most participants contribute little work, and that the bulk of the work is done by a small number of participants (Sauermann and Franzoni 2015; Sturn et al. 2015). This leads to relying on a few participants to disproportionately carry out the majority of the work, makes it less likely that projects will find such participants and be successful, and limits the amount of data that can be processed. Additionally, Munro et al. (2013) propose paid crowdsourcing as a cost-effective alternative when recruiting volunteers is not feasible; this may arise for less "prominent" disasters, which affect smaller areas, receive less media coverage and thus may attract significantly fewer volunteers. Such a comparison can be seen between two events in 2010: the Haiti earthquake, which received more than 3,000 news stories within the first 10 days, and the Pakistan floods, which received 320 broadcast news stories and 730 print news stories in the same timeframe (Brookings-Bern Project on Internal Displacement 2011).

Amazon Mechanical Turk (*MTurk*) is a widely popular online marketplace for crowdsourcing. MTurk allows *requesters* to post Human Intelligence Tasks (*HITs*) for *workers* to complete for payment. Although primarily intended for use as a paid crowdsourcing platform, recent work has shown that workers on MTurk are motivated by more than simply money (Kaufmann et al. 2011) and that aspects such as the framed meaningfulness of a task can impact measures of worker performance (Chandler and Kapelner 2013). Therefore, we considered MTurk as a means to recruit participants who may not be motivated purely by payment, but also voluntarily assisting in disaster response.

In order to gain insight into the motivations of workers on MTurk, and the interplay of paid versus volunteer work, in the context of disaster response, we ran a HIT on MTurk based on the MapMill project. MapMill is a citizen science mapping project that allows the uploading and subsequent rating of geotagged aerial images. Originally created in response to the Deepwater Horizon oil spill (Warren 2010), MapMill has subsequently been used in the aftermath of Hurricane Sandy, assessing damage and producing a heatmap of damage intended for use in directing relief efforts (Munro et al. 2013).

We developed a HIT that isolated the image rating portion of the MapMill project, which presented participants with a sequence of aerial photos of the Colorado flooding from 2013 and asked them to identify images containing damage. We posted a HIT on MTurk that used a *required work payment scheme*: workers were paid a fixed amount to rate *at least* some minimum required number of images; workers could then voluntarily continue rating images if they desired. Workers also completed a post-survey, which asked why they provided any additional ratings, in an open ended structure. We wished to examine the following research questions:

RQ1: What is the impact of requiring work versus not requiring work?
RQ2: What is the impact of providing feedback on progress?
RQ3: What different motivations do workers provide for their contributions in the project?

Within the HIT, we ran a multivariate experiment that varied two properties of the task. To examine the effects of required work, we varied the minimum amount of work required to receive payment: requiring no work, and requiring two different small amounts of work. To examine the effects of progress feedback, we varied whether workers saw a progress counter—toward their required amount of work and the total amount of work available. Since workers received the same fixed payment regardless of how many ratings they provided beyond the required minimum, we expected that any additional ratings would be provided for some reason other than payment, especially on the MTurk platform, where there are many other paying tasks available that might be a more lucrative use of a worker's time. Based on the *crowding-out effect* from psychology—where introducing an external reward can decrease intrinsic motivation and engagement (Lepper et al. 1973; Pretty and Seligman 1984)—we expected that paying for completing some work would generally have a negative impact on worker output and subjective experience. We also expected that showing workers their progress would have a positive impact on output.

After running the study, we observed that measures of productivity were generally improved by not requesting any work to be done at all, which is consistent with the crowding-out effect. Our observations were reinforced by the results of the open ended survey, which highlighted a switch in motivation from personal interest in the task to providing enough ratings, per the instructions. We further observed that the presence of progress feedback led to an overall increase both in amount of work and total time spent on the task.

This work contributes an empirical study to the growing understanding of crowdsourced worker motivations and behavior in disaster response. The study is consistent with motivational crowding-out, and the benefits of providing feedback on progress, in crowdsourced disaster response tasks. The open ended nature of the survey questionnaire about worker motivations allowed a wider range of elements to be identified, compared to standard multiple choice surveys, which proves encouraging for future research.

RELATED WORK

Crowdsourcing for Disaster Response

Crowdsourcing has featured prominently in disaster response scenarios as a means to help process the massive amounts of data that become available during these events. Efforts have been made in designing interfaces that will incorporate community-sourced intelligence into federal response operations (Crowley 2013). The evolution of crisis crowdsourcing led to the development of the Crisis Crowdsourcing Framework by Liu (2014), who also points to social, technological, organizational, and policy interfaces that need to be designed to guarantee an effective implementation of the framework. Goodchild and Glennon (2010) focus on the Santa Barbara wildfires that took place from 2007 to 2009 and how the community was able to contribute to disaster management through volunteered geographic information.

Aerial images are becoming increasingly important in disaster response. AIDR (Artificial Intelligence for Disaster Response), was initially designed to classify Twitter posts created during disasters using crowdsourcing (Imran et al. 2014). The platform was used to classify posts during the 2013 Pakistan earthquake. Ofli et al. (2016) extended AIDR to support aerial data captured via unmanned aerial vehicles (UAVs). Volunteer performance in rating aerial images was also examined in Mapmill, a community-sourced damage assessment project, following Hurricane Sandy. Munro et al. (2013) conclude that only four to six workers per image were required to ensure accuracy of the assessment, which can reduce the cost of aerial image assessment for future disasters. The use of Spatial Video Technology in damage assessment was also examined by Lue et al. (2014), where users reviewed video recordings of homes affected by the disaster. Inexperienced users were able to produce similar results to experts, suggesting the possibility of opening disaster response systems to a wider audience.

Finally, social media has been shown to be a potential source of information in disaster response operations. Tweets have been examined in assessing earthquake events near the islands of American Samoa and the city of Padang on Indonesia's island of Sumatra, in 2009 (Kireyev et al. 2009), as well as Hurricane Sandy (Kryvasheyeu et al.

Question	Choices
Can you tell us why did you complete the number of images that you did?	*(Free-response text area)*
Check all that apply:	*(Checkboxes to select any combination from:)* I did not understand the instructions. I thought I would get paid more. I thought my submission would not get approved.

Table 1. Survey questions and answers.

2016). Tapia et al. (2011) stress the need to overcome barriers in adopting microblogged data by international humanitarian relief organizations by introducing solutions that will ensure data reliability.

Crowd Motivation, Feedback, and Payment Schemes

Work in motivational psychology has established that *extrinsic* financial incentives can have a negative impact on *intrinsic* motivation (Deci 1971). This is commonly known as the "crowding-out" of intrinsic motivation (or "overjustification" effect). This effect has been explored empirically in many laboratory and field scenarios (Ryan and Deci 2000; Frey 1994). Other work has proposed that providing financial incentives may crowd-out intrinsic motivation to participate in research (Achtziger et al. 2015). Gneezy and Rustichini (2000) found that, when paying students a fixed amount to participate in an exam, adding a further small financial incentive reduced the student's exam performance relative to those who were not offered any further incentive; however, a larger financial incentive improved exam performance.

While tasks on MTurk involve payment, workers are motivated by factors other than money. Chandler and Kapelner (2013) found that framing an MTurk task in a more meaningful way led to an increase in participant engagement, as well as quantity of work, without a trade-off in quality. Kaufmann et al. (2011) concluded that although extrinsic motivation, in the form of payment, is present for MTurk workers, intrinsic motivation, such as skill variety, task identity, task autonomy, as well as community based motivation, is also significant.

However, Ho et al. (2015) found that performance-based payments may also improve quality, depending on the task. They argue that most MTurk tasks are *implicitly* performance-based, as workers consider that their work may be rejected if their performance is poor. Workers may come to a task with different notions of what is acceptable performance to avoid rejection. DellaVigna and Pope (2016) ran a large-scale study on MTurk, examining payment incentives and motivation, and found that even a very low payment did not notably crowd-out motivation; similar to other research on MTurk, they found that increasing piece-wise payment increased the quantity of work done. However, DellaVigna and Pope's task was quickly alternating keypresses, and thus may not be directly applicable to motivations in more "meaningful" tasks such as disaster response.

Previous crowdsourcing work has used a payment scheme with fixed payment for a set amount of required work. Khajah et al. (2016) paid workers a fixed amount to play a game for a minimum amount of time and then measured engagement as the amount of time played beyond that point, while Cai et al. (2016) paid workers a fixed amount to complete a small number of writing tasks.

Additionally, providing *feedback* on work is generally considered to improve worker motivation, as in the Job Characteristics Model of Hackman and Oldham (1976). Research shows that providing feedback on accuracy can improve the accuracy and quality of crowdsourced work (Riccardi et al. 2013; Dow et al. 2012).

Another form of feedback, *progress feedback*, gives workers feedback on the amount of work they have done towards a goal, without giving feedback on their accuracy. This can be particularly useful, as it does not require knowing ground truth for any tasks. In practice, this type of feedback may be more closely connected in Hackman and Oldham's model (1976) to *task identity* rather than what they term *feedback*, as it does not inform workers of their accuracy, but may give them a greater sense of contributing to a "whole" piece of work by filling the progress count.

Other work has examined the use of progress feedback in a crowdsourced setting. Toomim et al. (2011) found that workers preferred an interface that included text informing them how many more CAPTCHAs they had to complete to finish a HIT, but this feedback was combined with numerous other *aesthetic* changes in their comparisons. Chandler and Horton (2011) used artificial progress bars to indicate how close an image was to reaching its desired number of labels, and found that the positioning and balance of progress bars could be used to influence which

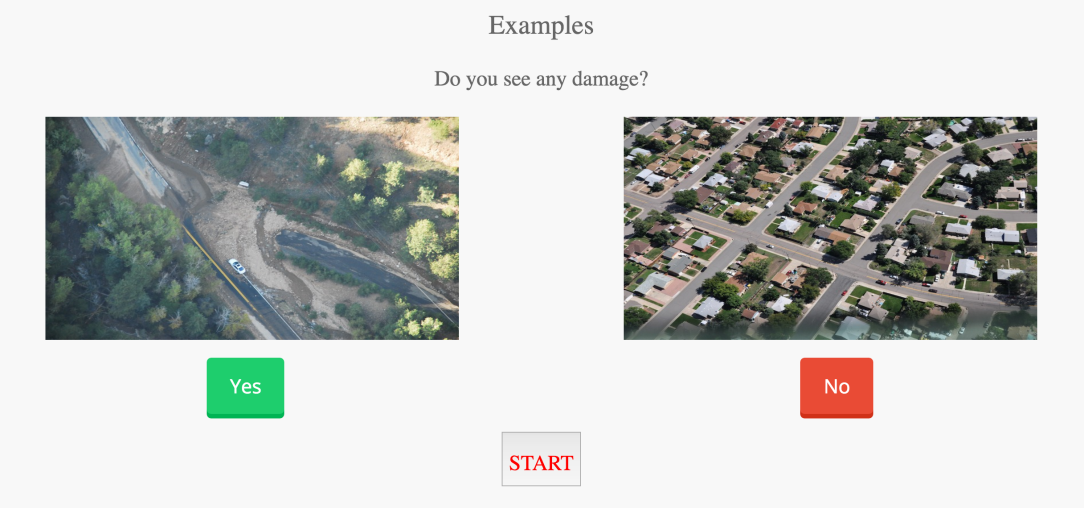

Figure 2. Screenshot of the training page used for our task.

image workers would provide labels for—generally increasing labeling for images that appeared to need more labels. Jacques and Kristensson (2013) found that the inclusion of a progress bar along with other "value proposition" elements (such as an example task) increased the conversion rate of workers who complete a HIT compared to those who are exposed to it.

TASK DESIGN

The motivation behind our task was to design a simple, user-friendly, and lightweight tool—inspired by the MapMill project—for participants to provide ratings for images, selected from a predefined set of possible ratings, to answer specific questions. We chose an image labeling task, as these are a highly popular and flexible method of crowdsourcing that requires less effort than other more complex tasks and can be used in a variety of settings. Using geotagged images of an area also allows immediate translation of responses for mapping purposes.

We posted a HIT on MTurk, which paid 50¢, titled:

> Disaster Area Map Image Labeling

The HIT description was:

> Label and categorize aerial images of disaster areas and answer a short 5 question survey on task.

The HIT keywords were:

> image; label; labeling; categorize; map

Because of the nature of the experiment, IRB approval was received. Upon accepting the HIT, subjects were informed that they were taking part in an experiment and were required to consent before proceeding. They were then provided with instructions on rating the images that appeared onscreen, including the minimum number of ratings (or *subtasks*) required. In order to counteract the variety of subjective beliefs about the amount of work required to be approved, the instructions explicitly stated that the submission would be approved if the survey was completed—similar to the "guaranteed payment" of Ho et al. (2015). The instructions were shown exactly the same to all workers except for the *conditional element*, which varied based on their experimental condition (described in more detail below). The instructions were:

> Click **START** to begin rating images. For each image, answer the question that appears beneath the image. When you are done rating, click the Go To Survey button to complete a short survey, after which your HIT will be completed. Please do not use your browser's back button while rating images or taking the survey. **If you complete the survey, your submission will be approved. This set contains 250 subtasks.** *[Conditional element of instructions, describing how much work was required, appeared here.]*

After an example for the types of rating, which acted as a small tutorial (shown in Figure 2), they then proceeded to rate images by answering damage related questions. When the required amount of work was completed, participants could either continue rating images or choose to complete the task by answering a short survey. If a participant rated all possible images, they were automatically taken to the survey.

Figure 3. Comparison of differences in task presentation among conditions.

In the survey, participants were asked to provide some feedback about the reasons that motivated their performance. The survey included two questions about motivation: the first question was a free-response question about the amount of work they did, and the second question included checkboxes allowing workers to indicate what we expected might be points of confusion about the instructions or payment scheme. A summary of the survey questions can be found in Table 1.

The data set used for this study contained aerial imagery publicly available through the Hazards Data Distribution System (HDDS), provided by the U.S. Geological Survey (Hazards Data Distribution System Explorer 2016). In particular, aerial imagery captured by Civil Air Patrol during the Colorado Floods that took place in 2013 was chosen to be rated by participants of the study.

Participants were asked to rate up to 250 images randomly chosen from the Colorado floods data set. They were presented with a simple interface, containing an image, the question "Do you see any damage?" and buttons for three possible answers, "Yes", "Maybe", and "No". They were also provided with a reminder of the instructions, "This set contains 250 subtasks. *[Conditional instructions.]*". The layout of the image rating page can be seen in Figure 1.

EXPERIMENT DESIGN

We carried out a 3x2 between subjects experiment design, using the following factors and levels.

- Amount of `required work`:
  - `0+`: Participants could rate as many images as they wanted (even none) and finish rating at any time by proceeding to the survey. This examined requiring no work.
    - The *conditional instructions* were: "**You may complete any number of subtasks before clicking the "Go to Survey" button to finish.**"
    - The "Go to Survey" button was always present.
  - `1+`: Participants had to rate at least one image before proceeding to the survey. This examined requiring the smallest amount of work possible, to see if this would crowd-out motivation.
    - The *conditional instructions* were: "**You must complete at least 1 subtask before clicking the "Go to Survey" button to finish.**"
    - The "Go to Survey" button appeared after 1 image was rated.
  - `10+`: Participants had to rate at least 10 images before proceeding to the survey. This examined requiring an order of magnitude more more work than 1+, but still a relatively small amount.
    - The *conditional instructions* were: "**You must complete at least 10 subtasks before clicking the "Go to Survey" button to finish.**"
    - The "Go to Survey" button appeared after 10 images were rated.

Variable	Description
Image Count	The total number of images rated.
Extra Image Count	The number of images rated beyond those required.
Total Time	Total time spent rating images, in seconds; the time between seeing the first image and moving on to the survey. We identified and excluded breaks of more than 5 minutes when rating an image.
More than Required	Whether the worker rated more images than the required amount.
Finished	Whether the worker rated the whole set.
Abandoned	Whether the worker accepted the HIT but abandoned it before completing the survey. Note that for this variable, all workers who accepted the HIT were included.
Agreement	Agreement with consensus. As ground truth ratings for the images were not previously known, we calculated agreement as the percentage of images for which a worker selected the consensus rating.

Table 2. Summary of the *performance* variables, based on workers' performance during the rating part of task.

Variable	Description
Checked-Understand	Whether the "I did not understand the instructions" box was checked.
Checked-Paid-More	Whether the "I thought I would get paid more" box was checked.
Checked-Rejected	Whether the "I thought my submission would not get approved" box was checked.

Table 3. Summary of the *survey checkbox* variables, based on workers' selection of the survey checkboxes.

- Presence of `progress feedback`:
  - `N`: No feedback on the current progress of ratings provided was present.
    - No information about the number of images rated was shown.
  - `P`: Feedback was given to the participants in the form of progress counts, showing how many images they have rated, as well as how many ratings were required (if any) and the total size of the set. As a design decision for simplicity in potentially showing progress towards two targets (completing required work and completing the entire set), we showed a textual representation of progress count rather than a graphical bar.
    - The text "**Progress: *L* / 250**" was shown.
    - The text "Required: *L* / *R*" was shown until the "Go to Survey" button appeared.

    (Where *L* is the number of images rated and *R* is the number of ratings required to go to the survey.)

Other than the variations described here, workers received identical tasks. This resulted in 6 conditions, which we refer to using `0+N`, `1+N`, `10+N`, `0+P`, `1+P`, `10+P`. Screenshots demonstrating the differences in conditions can be seen in Figure 3.

RESULTS AND ANALYSIS

We recruited 602 MTurk workers who completed the HIT, with an additional 61 workers who started—but did not complete—the HIT. Workers were assigned randomly into the 6 conditions. Workers were paid a flat rate of 50¢ if they completed the survey, regardless of the amount of work they did beyond the required limit.

Our analysis focused on three types of variables: *performance*, *survey checkbox*, and *survey free-response* variables. Performance variables were based on worker actions logged during the image rating portion of the HIT. Survey free-response variables consist of categories—identified using an open coding scheme as discussed below—from the free-response motivation question, treating each category as a separate Boolean variable. Survey checkbox variables are also treated as separate Boolean variables. A summary of all variables and their definitions is given in Tables 2, 3 and 4.

To analyze workers' responses to the free-response survey question, an open coding approach was followed. We began with the pre-defined categories *N/A*, *UNDERSTAND*, *PAID-MORE*, and *REJECTED*, to cover unrelated responses and loosely correspond to the options in the checkbox question. Two annotators where tasked with independently going over a sample of the submitted responses (the first 100) and identifying main categories. The two independent sets of categories were then used as a basis for a common coding scheme for the whole set of workers'

Variable/Category	Guideline	Example Response
DISASTER	Were interested in disasters.	*"Live in tornado alley. Interesting to see if anything looked like tornado damage."*
DO-MORE	Wanted to do as much as possible or finish the whole set. Also wanted to be thorough or do more than expected.	*"Wanted to finish the whole thing"*
ENJOY	Enjoyed the task, found it interesting and fun.	*"It was fun and interesting."*
EXTERNAL	Mentioned external obligations that did not allow them to continue, such as going to work.	*"Short on time and had to get to work."*
FEELING	Felt that what they did was the right amount or, wanted to do a certain number or, felt ready to move on to survey.	*"I thought I had reached 50 images submitted."*
HELP	Wanted to help the project.	*"I tried to complete as many as possible to contribute to the study."*
INSTRUCT	Said they did what the instructions said or the minimum.	*"because it asked to do at least 10 images"*
LOST-INTEREST	Mentioned getting bored or the task was becoming repetitive.	*"Was enjoying the task and stopped when it became monotonous"*
LOST-TRACK	Lost count of how many images they had rated or forgot to go to the survey.	*"wasn't thinking of the amount until I saw the survey button"*
PAID-MORE	Thought they would get paid more.	*"I completed 10 tasks looking for a bonus."*
REJECTED	Thought their submission would get rejected or not approved, wanted to make sure they got paid.	*"I tried to complete enough images to ensure i will get paid"*
SEE-MORE	Wanted to see more images.	*"Wanted to look at more"*
SKILL	Wanted to get better at the task, or thought they were good at it. Also concerned about accuracy of their work.	*"I felt that I got the general gist of the types of imagery I would see."*
TECH	Mentioned technical reasons.	*"I completed the number of images I did because the software had a slight delay for each image that was loaded."*
UNDERSTAND	Didn't understand the instructions or how many they were supposed to do. Thought the instructions were unclear.	*"I was unclear if I needed to do them all or if just by looking at one I could go on to the survey."*
VALUE	Considered the amount they did appropriate for payment, wanted to do enough work for how much they were getting paid.	*"Well honestly as much as I enjoy aerial images and high res airborne imagery there is only a certain amount of time I'm going to spend for $0.50."*
OTHER	Other reasons.	*"Because I felt the survey would be indepth and the focus of this HIT."*
N/A	Blank responses, numbers, not addressing the question.	*"Observed"*

Table 4. Summary of the categories for workers' text responses, along with exemplar responses, as a result of applying an open coding scheme. Each category also corresponds to a *survey free-response* variable.

responses. The responses were then again independently categorized, based on the agreed categories—during this second pass, the *DO-MORE* category was introduced. The final categorization was produced after the resolution of disagreements. The annotators remained the same throughout the various stages of the process. Inter-annotator agreement before resolution was measured using Cohen's kappa coefficient, with a resulting $\kappa = 0.67$, which is described by Landis and Koch (1977) as "substantial" agreement and Fleiss et al. (2013) as "fair to good" agreement. In total, 18 different categories were identified, which are summarized in Table 4.

For our statistical analysis, we performed omnibus tests for each variable to identify significant differences in `required work` and `progress feedback`, and identify possible interactions. For significant omnibus tests, we followed with post-hoc pairwise tests. As numerical variables were not normally distributed (determined using a Shapiro-Wilk test), we used Aligned Rank Transform (ART) for omnibus tests, the Wilcoxon rank-sum test for post-hoc tests, rank-biserial correlation (r) to compute effect sizes, and we report medians. ART (Wobbrock et al. 2011) is nonparametric and suitable for handling multiple factors and also allows us to analyze interaction effects. The ART procedure transforms data so that a factorial ANOVA can be applied. With Boolean variables, we used logistic regression for omnibus tests, Pearson's chi-squared test for post-hoc tests, phi (ϕ) to compute effect sizes, and we report percentages.

If no interaction was present in the omnibus test, with post-hoc tests we tested main effects present for `required work` (i.e. `0+` ~ `1+`, `0+` ~ `10+` and `1+` ~ `10+`) with a Bonferonni correction of 3 and `progress feedback` (i.e. `N` ~ `P`) with no correction. If an interaction was present, we performed post-hoc tests both within each feedback

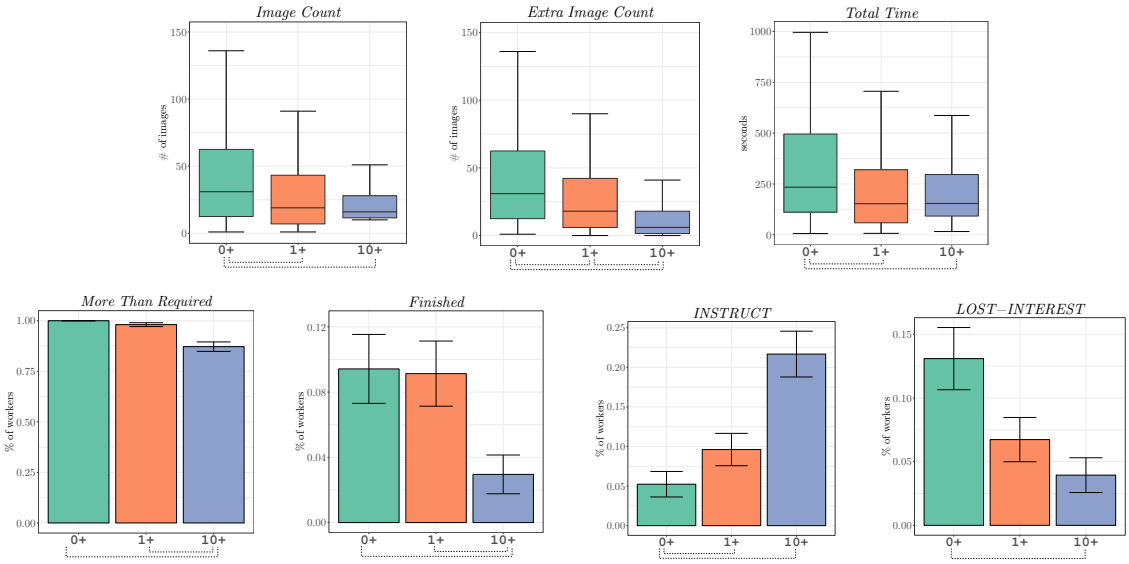

Figure 4. Summary plots of statistically significant post-hoc comparisons for `required work` main effects. For numerical variables, 25th, 50th (median) and 75th percentiles are shown, along with minimum and maximum values (whiskers), excluding outliers. For Boolean variables, percentages along with standard error bars are shown. Dashed lines indicate significant comparisons.

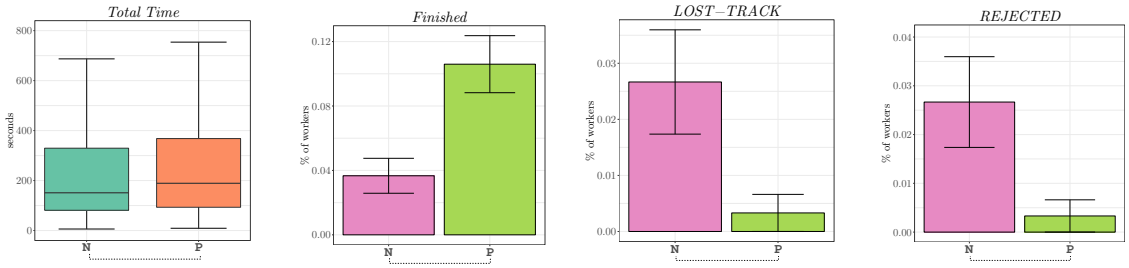

Figure 5. Summary plots of statistically significant post-hoc comparisons for `progress feedback` main effects. For numerical variables, 25th, 50th (median) and 75th percentiles are shown, along with minimum and maximum values (whiskers), excluding outliers. For Boolean variables, percentages along with standard error bars are shown. Dashed lines indicate significant comparisons.

condition and between (i.e., `0+N` ~ `1+N`, `0+N` ~ `10+N`, `1+N` ~ `10+N`, `0+P` ~ `1+P`, `0+P` ~ `10+P`, `1+P` ~ `10+P` `0+N` ~ `0+P`, `1+N` ~ `1+P` and `10+N` ~ `10+P`) with a Bonferonni correction of 9.

Tables 5, 6 and 7, along with Figures 4 and 5 contain a summary of the variables that were found to have statistically significant differences. As the post-hoc comparisons of interactions for *Checked-Paid-More* were not significant, we omit their inclusion from some summary figures and tables. We observed no significant variations among conditions in the *Agreement* variable (86.6% overall) and the *Abandoned* variable (9.2% overall) and therefore, we have omitted them from further analysis.

The vast majority of workers did at least some more work than was required: 572, or 95.0%, of workers, across all conditions, provided additional ratings. The number of ratings provided increased when no specific amount was required, versus requiring a small amount; the median amount of ratings offered in the `0+` conditions was 31, when compared to 19 for the 1+ conditions ($p < .001$). Moreover, the presence of feedback in the form of progress counts resulted in more workers completing the entire set. We observed significantly more people finishing the set in the `P` conditions (10.6%) than in the `N` conditions (3.7%, $p = .02$) Worker retention was the highest in cases with progress counts, with 10.6% of workers in the conditions with progress feedback finishing the entire set of 250 images. Requiring a small amount of work led to a sharp drop in continuing workers, for example, shortly after the lower bound of 10 images was reached. This can be seen in the chart of worker retention presented in Figure 6a.

Based on the survey results, workers whose responses indicated a loss of interest (*LOST-INTEREST*) were significantly more in the `0+` conditions, as compared to the `10+` conditions. This indication of interest points to

Variable	RWrk	Fdbk	RWrk × Fdbk	0+N	1+N	10+N	0+P	1+P	10+P
Workers				90	108	102	101	100	101
Image Count	$p < .001$			29	15	17	33	21	15
Extra Image Count	$p < .001$			29	14	7	33	29	5
Total Time	$p < .001$	$p = .011$		201s	127s	148s	252s	188s	156s
More than Required	$p < .001$			100.0%	96.3%	85.3%	100.0%	100.0%	89.1%
Finished	$p = .001$	$p < .001$		4.4%	5.6%	0.9%	13.9%	13.0%	4.9%
Checked-Paid-More			$p = .04$	23.3%	26.0%	13.7%	13.9%	27.0%	23.8%
INSTRUCT	$p < .001$			6.7%	9.3%	19.6%	3.9%	10.0%	23.8%
LOST-INTEREST	$p = .002$			15.6%	9.3%	4.9%	10.9%	4.0%	2.9%
LOST-TRACK		$p = .011$		1.1%	2.8%	3.9%	0.9%	0.0%	0.0%
REJECTED		$p = .011$		1.1%	5.6%	0.9%	0.9%	0.0%	0.0%

Table 5. Summary of results for variables with statistically significant differences from omnibus tests. Values are based on workers who completed the HIT; worker counts for each condition are given in the top row). Numerical variables are given as medians and Boolean variables as percentages. Numerical variables were tested with the Aligned Rank Transform and Boolean variables with logistic regression.

workers being more motivated by their own intrinsic motivation—up to the point where they decided to finish the task—when no work was required. We observed a high proportion of workers reporting that the primary reason for their performance was a result of following the instructions (*INSTRUCT*) in the 10+ conditions, in particular, 21.7% of workers. This was significantly higher than in the 0+ and 1+ conditions. This may indicate that, when more work was required, workers were motivated more by simply complying with the instructions to receive their extrinsic reward payment.

Levels of *LOST-TRACK* and *REJECTED* category were low across all conditions. However, we found both to be significantly lower in the P conditions, indicating that the progress bar may have given workers some additional clarity on how they were performing.

Looking at responses to the checkbox survey question, we found no significant differences in the post-hoc tests, indicating that among conditions, workers had a similar level of understanding of the instructions and payment scheme. Across all conditions, 9.2% of workers *Abandoned* the HIT (relative to all who initially accepted it), which was not significantly different between conditions. Results also indicate a high percentage of *Agreement*, 86.6% overall, which did not significantly vary between conditions. Thus we did not see an indication that worker accuracy was affected.

We did observe some minor confusion regarding instructions (*UNDERSTAND*, 3.9% of free responses) and payment (*PAID-MORE*, 0.2% of free responses). However, these were not different across conditions. *Checked-Paid-More* was found to have a significant interaction in the omnibus test, but not in the post-hoc tests.

Regarding our research questions, in the context of paid crowdsourcing for disaster response in an image rating task:

RQ1: What is the impact of requiring work versus not requiring work?
◇ When less work was required, workers did more work, were more likely to indicate stopping due to a loss of interest, and were less likely to indicate stopping due to the instructions.

Workers in the 0+ conditions did significantly more work beyond their requirements than the 1+ and 10+ conditions (*Extra Image Count*), and overall provided significantly more ratings (*Image Count*), and spent more time on the work (*Total Time*). On the contrary, workers in the 10+ conditions were significantly less likely to provide more work than required than workers in lower requirement conditions (*More than Required*), as well as less likely to complete the entire set (*Finished*).

Requiring work also impacted self-reported worker motivation. Analyzing workers' reported motivations for their performance, we observed that workers in the 10+ conditions were more likely to report complying with instructions as the primary reason for providing the amount of ratings they did, with 21.7% of the responses in these conditions belonging to the *INSTRUCT* category. In contrast, workers in the 0+ conditions were more likely to express interest up to the point when they decided to finish the task, with 13.1% of the responses in these conditions belonging to

Variable	`0+ ~ 1+`	`0+ ~ 10+`	`1+ ~ 10+`
Image Count	31 ~ 19 $p < .001, r = 0.23$	31 ~ 16 $p < .001, r = 0.25$	
Extra Image Count	31 ~ 18 $p < .001, r = 0.27$	31 ~ 6 $p < .001, r = 0.56$	18 ~ 6 $p < .001, r = 0.33$
Total Time	234s ~ 153s $p = .002, r = 0.20$	234s ~ 154s $p = .005, r = 0.18$	
More than Required		100.0% ~ 87.2% $p < .001, \phi = 0.25$	98.1% ~ 87.2% $p < .001, \phi = 0.20$
Finished		9.4% ~ 3.0% $p = .04, \phi = 0.13$	9.1% ~ 3.0% $p = .047, \phi = 0.12$
INSTRUCT		5.2% ~ 21.7% $p < .001, \phi = 0.23$	9.6% ~ 21.7% $p = .004, \phi = 0.16$
LOST-INTEREST		13.1% ~ 3.9% $p = .006, \phi = 0.16$	

Table 6. Summary of statistically significant post-hoc comparisons for `required work` main effects. *Finished* had no such comparisons. Numerical variables are given as medians and Boolean variables as percentages. Numerical variables were tested with the Wilcoxon rank-sum test and Boolean variables with the chi-squared test, applying the Bonferroni correction.

Variable	`N ~ P`
Total Time	151s ~ 189s $p = .019, r = 0.11$
Finished	3.7% ~ 10.6% $p = .020, \phi = 0.13$
LOST-TRACK	2.7% ~ 0.3% $p = .043, \phi = 0.18$
REJECTED	2.7% ~ 0.3% $p = .043, \phi = 0.18$

Table 7. Summary of statistically significant post-hoc comparisons for `progress feedback` main effects. *Total Time* had no such comparisons. Numerical variables are given as medians and Boolean variables as percentages. Numerical variables were tested with the Wilcoxon rank-sum test and Boolean variables with the chi-squared test, applying the Bonferroni correction.

the *LOST-INTEREST* category. Compared with the previous quantitative measures, these results lend support to the presence of the motivation *crowding-out* effect.

RQ2: What is the impact of providing feedback on progress?
◇ Providing progress feedback resulted in more work done and improved perceived clarity of the task.

Workers in conditions with progress feedback were significantly more likely to complete the entire image set of 250 images (*Finished*), as well as spend more time on the task (*Total Time*). These results indicate that providing progress feedback increased worker engagement in the task.

Giving workers progress feedback also reduced the amount of losing track of what they had done (*LOST-TRACK*) and fear of work being rejected (*REJECTED*).

RQ3: What different motivations do workers provide for their contributions in the project?
◇ Workers provided a variety of motivations other than payment for their participation.

Our open coding approach revealed a variety of different motivations for workers performing beyond their requirements. Figure 6b depicts the breakdown of categories in responses to the first survey question.

The task was well received by workers, with 10.7% of the workers' free-response answers falling into the *ENJOY* category and another 4.5% clearly stating wanting to help the project as a primary reason for rating more images. Another 8.5% of workers reported that they felt the amount they had reached was adequate to their opinion and felt ready to finish the HIT (*FEELING*). The above responses indicate intrinsic motivations regarding the worker's

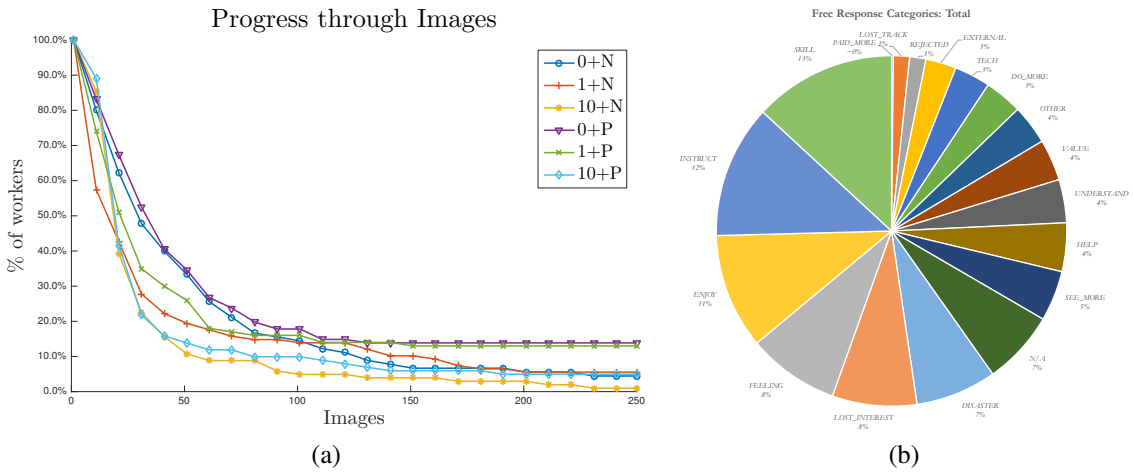

Figure 6. (a) Worker retention chart by image count. The x-axis shows the image count and the y-axis shows the percentage of workers in a condition that rated at least that many images. The rapid falloff after completing work in the 10+ conditions, and the large percentage of workers in the 0+P and 1+P conditions who competed all available images, can be clearly seen. (b) Breakdown of response categories for responses to the free-response motivational survey question.

personality and abilities. We also identified another motivation for workers, which has to do with the nature of the task, as 7.5% of responses fell into the *DISASTER* category. This group of workers were interested in analyzing disasters.

Notably, there were still some responses indicating potential misinterpretation of the instructions, including workers concerned about their work being rejected, even though we explicitly stated work would not be rejected, or expecting more payment for more work even though none was offered. This may be due to worker expectation based on the MTurk platform. This would indicate that clarity of instructions and setting clear expectations is of great importance.

CONCLUSION AND FUTURE WORK

In this work, we invited MTurk workers to participate in a disaster response scenario by asking them to provide ratings for images taken during the Colorado floods of 2013. We awarded a fixed amount of payment and varied the minimum amount of ratings required as well as whether or not progress feedback was provided. Our results indicate that most of the workers did more work than required to receive payment. We also observed that workers generally did more work due to their own interest when no minimum work requirement was asked of them.

Existing literature points to many other possible payment schemes, such as payment per unit of work completed, as well as bonuses per goal accomplished (Ho et al. 2015). Future work could compare these per-unit schemes to a required work scheme. Moreover, we explored 3 different levels of required work for the purposes of this study, with 10 images being the highest amount. We would like to examine the effects of requiring even larger amounts of work in participation and engagement in disaster response crowdsourcing.

One interpretation of giving workers flexibility in the amount of work they do is that it effectively allows them to set their own wage. In a survey of workers performed by Munro et al. (2013), they found a suggested wage of 0.1¢ to 2¢ per "judgment". In our work, we found that the effective wage in the 0+ conditions came to 0.9¢ per rating on average, which falls into that range (in contrast with the higher per-rating wages of 1.1¢ in the 1+ and 1.5¢ in the 10+ conditions.

Making crowdsourced tasks more interesting and engaging is a promising area for future exploration. In this work, we found that workers were more likely to indicate reasons related to interest for their participation when no work was required of them. Therefore, we believe that designing tasks that are more engaging and interesting for participants will have a bigger impact if a payment model is chosen where no work is required. This work is part of an initial effort into creating a platform that can be used to develop techniques for improving the experience of contributing to crowdsourced disaster response—for both paid crowd workers and volunteers.

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

Achtziger, A., Alós-Ferrer, C., Hügelschäfer, S., and Steinhauser, M. (2015). "Higher incentives can impair performance: neural evidence on reinforcement and rationality". In: *Social Cognitive and Affective Neuroscience* 10.11, pp. 1477–1483.

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.

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.

Brookings-Bern Project on Internal Displacement. (2011). *A Year of Living Dangerously: A Review of Natural Disasters in 2010*. http://www.refworld.org/docid/4dabde142.html.

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.

Chandler, D. and Horton, J. J. (2011). "Labor allocation in paid crowdsourcing: experimental evidence on positioning, nudges and prices". In: *Human Computation* 11, pp. 14–19.

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*.

Deci, E. L. (1971). "Effects of externally mediated rewards on intrinsic motivation". In: *Journal of Personality and Social Psychology* 18.1, pp. 105–115.

DellaVigna, S. and Pope, D. (2016). *What motivates effort? Evidence and expert forecasts*. Working Paper 22193. National Bureau of Economic Research.

Dow, S., Kulkarni, A., Klemmer, S., and Hartmann, B. (2012). "Shepherding the crowd yields better work". In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pp. 1013–1022.

Fleiss, J. L., Levin, B., and Paik, M. C. (2013). *Statistical Methods for Rates and Proportions*. John Wiley & Sons.

Frey, B. S. (1994). "How intrinsic motivation is crowded out and in". In: *Rationality and Society* 6.3, pp. 334–352.

Global Xplorer (2016). http://www.globalxplorer.org/.

Gneezy, U. and Rustichini, A. (2000). "Pay enough or don't pay at all". In: *The Quarterly Journal of Economics* 115.3, pp. 791–810.

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.

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.

Jacques, J. T. and Kristensson, P. O. (2013). "Crowdsourcing a HIT: measuring workers' pre-interactions on microtask markets". In: *Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing*.

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*.

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.

Kireyev, K., Palen, L., and Anderson, K. (2009). "Applications of topics models to analysis of disaster-related twitter data". In: *NIPS Workshop on Applications for Topic Models: Text and Beyond*. Vol. 1.

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Hentenryck, P. V., Fowler, J., and Cebrian, M. (2016). "Rapid assessment of disaster damage using social media activity". In: *Science Advances* 2.3, e1500779.

Landis, J. R. and Koch, G. G. (1977). "The measurement of observer agreement for categorical data". In: *Biometrics* 33.1, pp. 159–174.

Lepper, M. R., Greene, D., and Nisbett, R. E. (1973). "Undermining children's intrinsic interest with extrinsic reward: a test of the "overjustification" hypothesis". In: *Journal of Personality and Social Psychology* 28.1, pp. 129–137.

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.

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.

Mitry, D., Peto, T., Hayat, S., Morgan, J. E., Khaw, K.-T., and Foster, P. J. (2013). "Crowdsourcing as a novel technique for retinal fundus photography classification: analysis of images in the EPIC Norfolk cohort on behalf of the UKBiobank Eye and Vision Consortium". In: *PLoS ONE* 8.8, e71154.

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*.

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.

Pretty, G. H. and Seligman, C. (1984). "Affect and the overjustification effect". In: *Journal of Personality and Social Psychology* 46.6, pp. 1241–1253.

Raddick, M. J., Bracey, G., Gay, P. L., Lintott, C. J., Murray, P., Schawinski, K., Szalay, A. S., and Vandenberg, J. (2010). "Galaxy Zoo: exploring the motivations of citizen science volunteers". In: *Astronomy Education Review* 9.1.

Riccardi, G., Ghosh, A., Chowdhury, S. A., and Bayer, A. O. (2013). "Motivational feedback in crowdsourcing: a case study in speech transcription". In: *Proceedings of the 14th Annual Conference of the International Speech Communication Association*, pp. 1111–1115.

Ryan, R. M. and Deci, E. L. (2000). "Intrinsic and extrinsic motivations: classic definitions and new directions". In: *Contemporary Educational Psychology* 25.1, pp. 54–67.

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.

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*.

Tapia, A. H., Bajpai, K., Jansen, B. J., Yen, J., and Giles, L. (2011). "Seeking the trustworthy tweet: can microblogged data fit the information needs of disaster response and humanitarian relief organizations". In: *Proceedings of the 8th International Conference on Information Systems for Crisis Response and Management*, pp. 1–10.

Toomim, M., Kriplean, T., Pörtner, C., and Landay, J. (2011). "Utility of human-computer interactions: toward a science of preference measurement". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2275–2284.

Warren, J. Y. (2010). "Grassroots mapping: tools for participatory and activist cartography". Thesis. Massachusetts Institute of Technology.

Wobbrock, J. O., Findlater, L., Gergle, D., and Higgins, J. J. (2011). "The Aligned Rank Transform for nonparametric factorial analyses using only ANOVA procedures". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 143–146.