And Now for Something Completely Different: Improving Crowdsourcing Workflows with Micro-Diversions

Peng Dai [1], **Jeffrey M. Rzeszotarski**[2], **Praveen Paritosh**[1], **Ed H. Chi**[1]
[1]Google, Inc., 1600 Amphitheater Pkwy, Mountain View, CA 94043, USA
[2]Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA
daipeng@cs.washington.com, jeffrz@cs.cmu.edu, pkp@google.com, chi@acm.org

ABSTRACT
Crowdsourcing has become a popular and indispensable component of many problem-solving pipelines in the research literature, with crowd workers often treated as computational resources that can reliably solve problems that computers have trouble with, such as image labeling/classification, natural language processing, or document writing. Yet, obviously crowd workers are human, and long sequences of the same monotonous tasks might intuitively reduce the amount of good quality work done by the workers. Here we propose an investigation into how we can use diversions containing small amounts of entertainment to improve crowd workers' experiences. We call these small period of entertainment "micro-diversions", which we hypothesize to provide timely relief to workers during long sequences of micro-tasks. We hope to improve productivity by retaining workers to work on our tasks longer and to either improve or retain the quality of work. We experimentally test micro-diversions on Amazon's Mechanical Turk, a large paid-crowdsourcing platform. We find that micro-diversions can significantly improve worker retention rate while retaining the same work quality.

INTRODUCTION
In crowdsourcing task markets such as Amazon's Mechanical Turk (MTurk), hundreds of thousands of workers take on tasks that are challenging for a computer program to solve automatically. For example, workers create corpora of translation data [5], identify interesting visual features in astronomy data [32], and help curate structured data in semantic entity databases [29]. Workers may even do creative or complex work such as writing news articles or generating novel artifacts with design constraints [27, 43].

In paid-crowdsourcing platforms, workers can often perform a large amount of tasks in a single, prolonged setting. Large batches of tasks done in sequence are efficient for workers due to their improving familiarity with the work, and besides, it can be very time-consuming for workers to search for new, interesting work. For work requesters, batched work might be beneficial as well, since workers might develop expertise that can be incorporated into judgments [15, 41]. Moreover, long-term, diligent workers have been shown to support more complex workflows and tasks [14, 26, 27].

However, human workers are quite different from computer programs in their performance characteristics—intuitively, they are not nearly as tireless or reliable. We know from research in other domains that fatigue, both physical and cognitive, can affect workers doing large batches of tasks, not only risking reduced well-being, but also creating lower quality, unreliable data as a result [31]. Imagine, for example, a task asking workers to tag one image. A worker can easily tag the image, and probably have no problem tagging a few more. Now imagine someone who has been working on the same type of tagging tasks during the past hour. Seeking variety and to reduce monotonous work, she might choose to look for another type of task.

In an opt-in setting such as MTurk, workers may choose to stop working, or to find other tasks that pose different, refreshing challenges. Indeed, past MTurk studies show that workers tend to switch tasks once a certain goal has been accomplished [20]. Other research also has demonstrated situations where worker engagement in volunteer crowdsourcing fades over time [33]. Clearly, to obtain good quality and stable results, work requesters want to retain experienced workers.

We are not aware of existing crowdsourcing research on interventions that would help retain good-quality workers. There are, however, great research on how to mitigate quality control issues in crowdsourcing systems using external checks such as CAPTCHA questions, gold standards, majority voting, and behavioral logging [3, 5, 28, 38]. Moreover, crowdsourcing quality control techniques generally assume a consistent model per worker over time [16,41]. Interestingly, some research on temporal probabilistic models of multiple annotators exists [17].

CSCW 2015, March 14–18, 2015, Vancouver, BC, Canada.
ACM 978-1-4503-2922-4/15/03.
http://dx.doi.org/10.1145/2675133.2675260

Here we propose a method to help retain good workers through diversions [1] in the form of small entertaining interruptions. Imagine the previous worker who has been tagging images for 10 minutes. Suddenly she might receive an entertaining diversion. Perhaps she is instead asked to watch a short, funny video if she wants, or is told that she has done a good job so far, together with a leaderboard of how well she is doing compared to other workers. Such small interruptions do indeed incur a cost in context switching as the worker must return to their task afterwards [34, 42]. Yet, the act of interrupting also functions as an entertaining time-off from the tasks, potentially refreshing the worker.

We suggest that these refreshing *micro-diversions* might offer benefits that outweigh the tradeoff of context switching, providing a net benefit to crowdsourcing owners. It is important to note that these micro-diversions are different than merely adding fun content to a workflow. They are intended to purposely break the flow of the worker, possibly reducing fatigue, improving alertness, and refreshing workers' cognitive resources.

RELATED WORK

There exists a growing body of research studying human factors and behavior in crowdsourcing markets.

First, improvements in basic task design leverages human behavior to produce better answers. For instance, tasks can be structured so that they resist cheating by requiring more effort, or require explanations to reinforce consideration [25]. Optimizations on design parameters such as the number of questions to ask and payment structure help produce better results with the same budget and time constraints [21]. Even the framing of a task can have a big impact: Chandler and Kapelner [8] demonstrated that workers' perceptions influence their performance, finding that workers who believe that they are working for a charity do a better job. Providing feedback during the tasks, such as strategy advice or good example work by other workers, also result in better quality work [19]. All of these studies take advantage of human aspect of human computation using a Human-Computer Interaction research approach. However, no research that we are aware of thus far has directly tried to improve worker experience through entertaining diversions in crowdsourcing systems.

Second, intuitively, long sequences of monotonous crowdsourcing tasks incite boredom. Unsurprisingly, the psychological and human factors communities have explored boredom and its impact on work. One source for this research is the military, which found itself concerned with fighter pilots sitting in cockpits for long hours and radar operators staring at screens for entire evenings. These vigilance tasks require operators to conduct actions and devote attention over a long period of time. As the task progresses, researchers have found physiological and cognitive effects that reduce performance [36]. Krueger reinforces these findings, demonstrating a linear decay in performance over work time given sustained, possibly sleepless work [31]. Somewhat surprisingly, some tasks cause poor performance not because they are overly taxing, but because they underutilize mental resources, leading to boredom. Pattyn et al. [35] find that underutilization is related to misdirection of attention resources or withdrawal. Carriere et al. [6] relate an underutilization of attention or mistaken redirection to boredom. Later research has identified boredom as a unique condition distinct from underarousal or apathy [7]. However, in crowdsourcing, workers are already free to take breaks anytime they want, and they are also free to switch to other tasks if they choose so, therefore it is possible that micro-diversions are not needed to improve the overall worker experience. Instead, viewed from the work requesters' perspective, our research goal is to retain productive workers and help them stay on the requesters' tasks.

Over long working periods, the opposite of underarousal experiences are flow experiences [10], but these are hard to predict and likely require matching task difficulty to each worker's expertise. Flow experiences, despite being beneficial, might also upset quality control measures when workers unexpectedly increase the quality of their responses. Though relevant, this is outside of the scope of our research goals currently.

Third, while past research shows that diversions can ameliorate boredom, however, their effect also greatly depends on context. Good diversions should be different from the normal, tiring tasks and do not demand an excess of attentional resources [11]. Good diversions may have more impact if they have physical and cognitive conditions that are congruent to the task [18]. For instance, a diversion for a computer operator might involve using computer for something different or stretching the muscles an operator uses for sitting and typing. The length of a diversion interruption is delicate; not too long to require a high context-switch cost, but long enough to refresh the worker. Generally, even small 3- to 30-second breaks can have a measurable impact on workers [18]. This is encouraging for crowdsourcing markets, as micro-diversions might be slipstreamed into workflows automatically.

Fourth, interrupting a workflow has associated costs. Generally speaking, workers might perform poorly after an interruption to their work [30, 34]. Interruptions during more complex tasks take longer time to resume [22]. Interruptions also may increase stress [34]. HCI literature suggest that workers often underestimate the costs of interruptions [23]. The costs of interruptions are not consistent either. For participants working on a search task, interrupting them while they were evaluating multiple results was much more costly than interrupting

[1] We use the term 'diversion' because work breaks might be both full rest breaks as well as free-time breaks, in which workers can choose an entertaining activity of their own. Instead we are proposing to prescribe specific entertaining diversions without worker choice.

them while they performed the search query [12, 13]. This has lead to the introduction of numerous intelligent systems and models for locating more optimal moments for interruptions. These systems work to find natural breakpoints in the task or support better task/group awareness [1, 2, 37].

These findings suggest that in a crowdsourcing micro-diversion system, the timing of the interruptions is critical so as to minimize their workflow disruption. Luckily, crowdsourcing workflows have natural breakpoints at the end of task units in a batch. Even better, outside of the crowdsourcing literature, interruptions have been found to improve the performance of workers completing simple tasks by increasing vigilance [39]. We hope to demonstrate that we can retain more of the productive workers in a crowdsourcing system, even though they are free to leave any time.

In summary, based on prior research, here we investigate the potential benefits of micro-diversions, hypothesizing that they outweigh the costs of interruptions at natural breakpoints in human computation tasks. As the literature suggests, on the one hand, micro-diversions might refresh workers and alleviate their boredom. On the other hand, micro-diversions might interrupt the workflow and disrupt working mental states and their flow experiences.

RESEARCH QUESTIONS

We aim to perform experimental studies to find out if micro-diversions have any impact on worker retention and performance in crowdsourcing. Here are the research questions we aim to investigate:

RQ1: Retention Whether and how much does having micro-diversions improve worker retention?

If workers are refreshed by the micro-diversions, then they are likely to stay and do more of the instances of the same task. If they feel that diversions are a distraction, they might actually leave earlier than a no-diversion control condition.

Typical tasks that exist within crowdsourcing platforms might not benefit from micro-diversions, since they typically have short duration, and workers are free to move on to other tasks quickly. However, there is always a switching cost to another task type, since workers would have to search and select a new task, read the instructions, and cognitively prepare for another task structure.

RQ2: Worker Performance In terms of time-on-task and accuracy, do micro-diversions distract workers and bring their performance down, or instead refresh them and either maintain or even boost their performance?

Performance might be measured by both total task-duration per unit work (including the time cost of the diversions) as well as the task accuracy. For our purposes, an increase in time and/or a decrease in accuracy would be considered lower worker performance. If workers do not experience boredom or fatigue, then diversions simply causes increase in total task time. On the other hand, if the workers are refreshed by the micro-diversions, then they might either work faster or maintain/improve their accuracy despite doing more work.

RQ3: Contextual Effects Are different task types affected by the micro-diversions differently? What types of micro-diversions are the most effective in promoting worker engagement?

We aim to study different task types across several domains as well as different micro-diversion types to see if only certain task-diversion combinations work to refresh workers.

RESEARCH METHOD

We conducted our experiments on the Amazon's Mechanical Turk (MTurk) platform, which is a popular platform for crowdsourcing. Tasks on MTurk are grouped into *batches*, which is a set of tasks owned by the same requester that share the same user interface. Workers are paid by a fixed per task rate defined by task requesters. Workers retain full control over which task batch to work on. If a worker likes one batch, she can do as many tasks from the same batch. Therefore, beyond worker task accuracy and speed, the number of tasks completed per worker per batch is a good signal of worker engagements.

Experimental Design

The study used a 3 tasks $\times$ 3 micro-diversions between-subjects design. The three different types of tasks are Image Identification (II), Wikipedia article Quality rating (WQ), and Freebase Entity merging (FE). There were three types of micro-diversions: No Diversions (ND), Game (GD), and Serialized Story (SD). The three task types are crossed with the three micro-diversion types in a between-subjects study design, preventing participants from completing tasks in multiple task and diversion combinations.

In each condition, we provided at least an hour of work for workers to complete in each batch. We also introduced gold standard data so that we could measure work accuracy.

Subjects

We solicited 30 unique workers per condition to stay and do as much as 1-2 hours of work as they could in their assigned task and diversion type combination.

Therefore, the experimental design gives us 3 task types $\times$3 diversion types $\times$30 workers = 270 unique worker submissions.

Payment

After some deliberations, we decided not to pay workers for time spent in the diversion, since this would have been somewhat awkward for the Game Diversion. However, a worker can quickly skip over a diversion by clicking to proceed to the next question as soon as the diversion UI loads (typically just a few seconds).

This means, for each task type, as long as two workers answer the same number of questions, regardless of how many diversions encountered, they will receive the same amount of monetary reward.

Dependent Variables: Metrics

To measure whether micro-diversions had an effect on workers, we evaluated workers' engagement through retention as well as their average answer time and accuracy.

For RQ1 on retention, we measured the amount of task work units completed by a worker. Retention can be thought of as survival over the multiple instances of a task. The reason we use retention as the main metric here stems from behavioral economics. We regard Mechanical Turk workers as somewhat rational decision makers who decide whether they want to do more work or not each time they see a new task [9]. Since workers opt-in to MTurk tasks, a worker can choose to keep doing tasks in series until her rate of compensation outweighs her calculation of the costs of doing work. We assert that worker boredom and fatigue are part of this cost/benefit calculation that workers make. For example, while a worker may be willing to do many iterations of a task that tags different images, after some time the generous pay may no longer outweigh her boredom, and so she leaves. Thus, if workers in one condition are more engaged than the others, they will stay longer. If more workers stayed to do more tasks in a certain condition, we suggest that that condition was more engaging.

For RQ2 on work performance, we measured the response time and answer accuracy as follows:

Answer response time is measured from the moment a task work unit is served to the worker to the time that the server received the answer for that work unit from the worker. To be fair to the baseline condition containing no diversions, the time a worker spends on a diversion is aggregated with the time she answers the previous question. For example, the answer time a worker spends on one question contains the time she spends both on the question and the coming diversion, if there is one.

Worker accuracy is measured by comparing the workers' answers to the gold standards. A single worker's accuracy is calculated as the percentage of accurate answers in all of her work units. We then aggregated and computed the average of the worker accuracies per task-diversion condition combination.

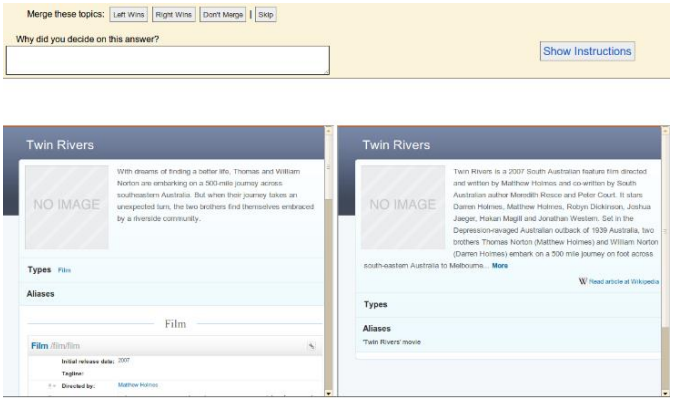
Independent Variables: 3 Task Designs

As depicted in Figure 1, we introduce three types of tasks here. Obviously, there are many type of tasks we could have studied, so we chose tasks that are somewhat representative of the types used in the research literature.

Wikipedia article Quality rating (WQ)

This task type is a cognitive-intensive task on reading a Wikipedia article and rating the quality of the article, similar to the task used in Kittur et al. [25]. Intended

(a) Wikipedia article Quality rating (WQ)



(b) Freebase Entity merging (FE)

(c) Image Identification (II)

Figure 1. Example task types from 3 task domains.

to represent a typical quality judgment task, we incorporated 50 articles from the Wikipedia featured articles list and 50 from the 'needs-cleanup' list that contains lower quality articles. This gave us a mix of good and bad articles that we could use as gold standard. On a Likert scale for article quality, we can measure worker accuracy by comparing whether a worker rated the featured articles better than the cleanup articles on average. As done in Kittur et al. [25], we also asked workers to provide a short summary of the page contents as a proof that they read the article as well as two questions concerning how many images and sections are in the article to deter cheaters. Workers could complete up to 100 tasks for 6 cents each.

Freebase Entity merging (FE)
This task type is also cognitive, making identity judgments for entity resolution of Freebase entities [4] (e.g. one entity might be a tennis player and another a movie star, both with the same first and last names.) In this task, we asked if two different entities were in fact the same entity, and if so, which node's content should get precedence in merging them (e.g. two celebrity entities, one being a nickname for the other.) Typical of a side-by-side judgment task, this task is highly cognitive, as the comparator has to examine both articles looking for consistencies and inconsistencies before making a decision on whether they are in fact referring to the same entity. To deter cheating we asked workers to cite evidence that supports their decision. We extracted 81 different 4-cent merge tasks using the Freebase logs of past entity resolution tasks, with the Freebase adjudicated decision as a gold standard.

Image Identification (II)
This task type is primarily perceptual, asking workers to determine whether each of 12 images contained scenic places or buildings (i.e. 12 images is one task unit.) Image classification and labeling tasks are often used in machine learning research. Compared to other cognitive tasks, perceptual tasks like this are short and relatively easy, so workers could do up to a batch of 100 different, 12-image tasks for 1 cent each. We sampled 600 of the images from panoramio that are of landscapes and landmarks. The other 600 images were sourced from public Google Picasa images, and human judges certified that they did not contain landscapes or landmarks.

Independent Variables: 3 Micro-Diversion Designs
Having outlined the task types above, we now discuss the design of the micro-diversions, which we debated amongst ourselves extensively. In particular, we debated whether (a) micro-diversions should simply be different from the real tasks, so that workers are refreshed just from doing something different? Or (b) diversions should be a real forced-timed rest break/interruption, even though that might risk having workers leave our task completely? Or (c) micro-diversions that are fun and engaging so that we might maximize worker retention and performance? Since there are many possible micro-diversions that we could examine, we decided to opt for (c), leaving the complexity of the issue for future work. Since we were not confident that micro-diversion actually improves retention and performance in crowdsourcing environments, we thought (c) is the most likely to produce an positive effect.

Here we introduce the three diversion types as shown in Figure 2:

Control / No Diversions (ND)
No Diversion is the baseline control condition.

Game Diversion (GD)
We used a simple dice game. Workers were presented with a dialog window that allows them to wager a proportion of the rewards they have accrued so far for more. For instance, the worker could spend 4 cents to have a 50/50 shot at 8 cents, or a 25% chance of getting 16 cents. After picking their wager or choosing not to wager anything, flashing graphics are shown about a rolling dice and finally the amount of payoff. This diversion employs the gambler's fallacy as a gamification mechanic. The gambler's fallacy relies around poor human odds estimation.

In our game, the odds were set completely *n*eutrally—that is, workers should get the same pay over the long run, whether they gambled or not. However, thanks to the gambler's fallacy, a few losses might encourage the worker to try again because they "feel a win coming".

Because the gaming process was easy and encouraged speed, this diversion type is somewhat perceptual, and might relate more easily to the perceptual image tagging (II) task rather than the highly cognitive entity merging (FE) task. Or alternatively, one might surmise that a perceptual diversion would be welcome after a long cognitive load period.

Story Diversion (SD)
Workers received one page of a Hugo-Award-winning narrative webcomic, Digger, by Ursula Vernon [40] (used with permission). This comic features compelling general-interest stories and rich visual designs that continue from page to page. For each diversion, a worker received one page of the comic in the sequence. Worker had to scroll to the end of the page to start their next task. By continuing to work one could see more of the comic and therefore learn more of the story. This introduces a motivation mechanic to continue, though utilizes some cognitive resources. The worker must think and read in order to appreciate the comic.

Because of its narrative and cognitive nature, this diversion might relate more closely to the Wikipedia (WQ) and entity (FE) tasks rather than the Image labeling (II) tasks. Alternatively, one might surmise that a cognitive reading task would be a welcome break to a long perceptual task.

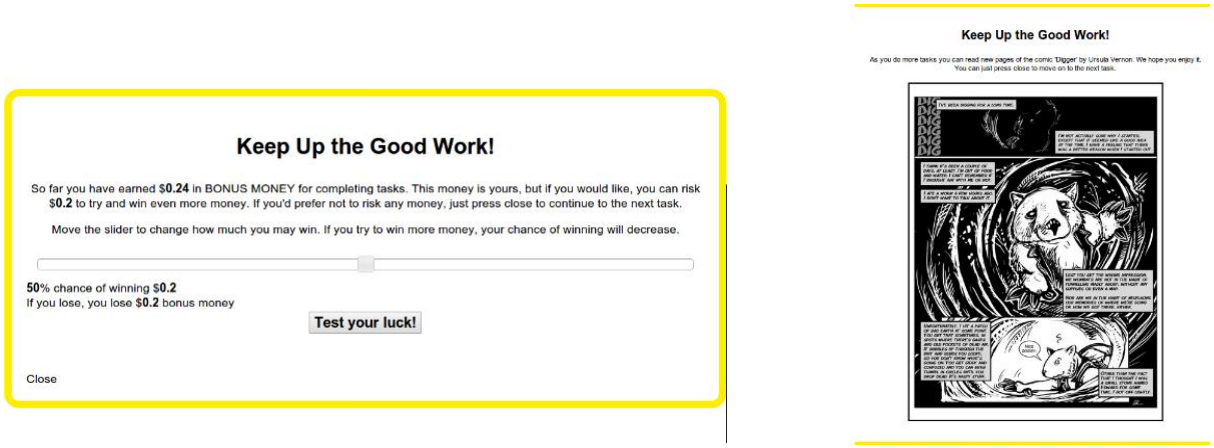


Figure 2. Example micro-diversions. The Control (No Diversions, ND) baseline is not shown here. Left: rolling the dice (Game Diversion, GD). Right: serialized comic (Story Diversion, SD).

Diversion delivery and experiment schedule

We wanted to interrupt the workers with a diversions approximately once every 10 minutes. Therefore, to determine the diversion schedule, we divided 10 minutes by the average answer time per question/work unit (called Human Intelligence Task / HIT on MTurk) and rounded the result to its nearest integer number f. To determine the average answer time per HIT, we ran a calibration pilot over approximately 30 workers for each task type. The values of f for the Wikipedia Quality (WQ), Freebase Entities (FE) and Image Identification (II) tasks were 4, 8, and 6 respectively. We adopted a static diversion schedule of every f HITs, so, on average, we expect a diversion to happen approximately every 10 minutes. We chose to deliver diversions after a set number of tasks rather than exactly 10 minutes to avoid interrupting workers mid-task.

As mentioned previously, a worker can quickly skip over a diversion by clicking to proceed to the next question as soon as the diversion UI loads (typically just a few seconds).

Experiment time: To ensure a similar participant pool, we posted tasks at the same time (4-5PM PST) each day and let them run until all tasks are completed or 48 hours have elapsed.

Subject Experimental Procedure

Server Software: We developed a web server on Google AppEngine that allowed us to deliver controlled batches of tasks to workers through Mechanical Turk. The server maintains state information on each worker.

The experimental procedure for a single worker is as follows:

When a worker asks for a task, her browser queries our server through the external question interface on MTurk. URL parameters encode metadata such as which conditions (task and diversion types) the worker is under. Upon receiving the request, the server examines its database for an entry for that worker's assigned condition combination.

If the worker is in the middle of the batch, it determines whether the worker is due for a diversion. If the worker is not due for a diversion, it randomly chooses a new, uncompleted task in the batch.

The worker is free to leave any time they wished. When they first start, workers are informed they are free to leave at any time and collect payment for the tasks completed so far. In the interface workers saw a large "Finish & Collect Payment" button at the bottom of the screen, which they could press to finish early, or by using standard MTurk UI. We designed our tasks so that Amazon recorded each work unit (HIT) that the workers completed. Later, we paid workers by referring to HIT counts.

The system closes out an experimental condition after a maximum of 30 workers have participated.

RESULTS AND ANALYSIS

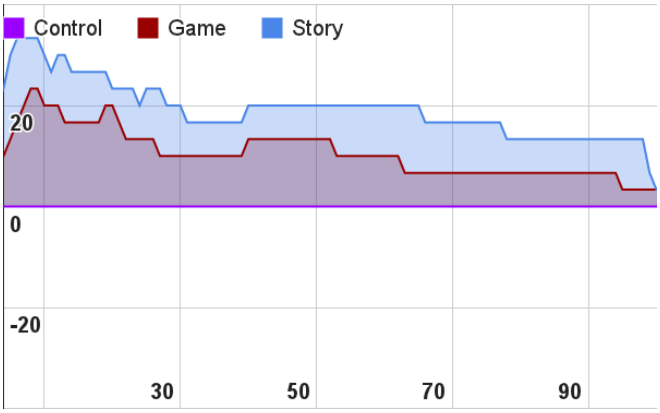
In this section, we present the results from our micro-diversion experiment. First, we perform detailed analysis on worker retention. Next, we focus on worker accuracy under various conditions, then we analyze the time on task.

RQ1: Worker Retention

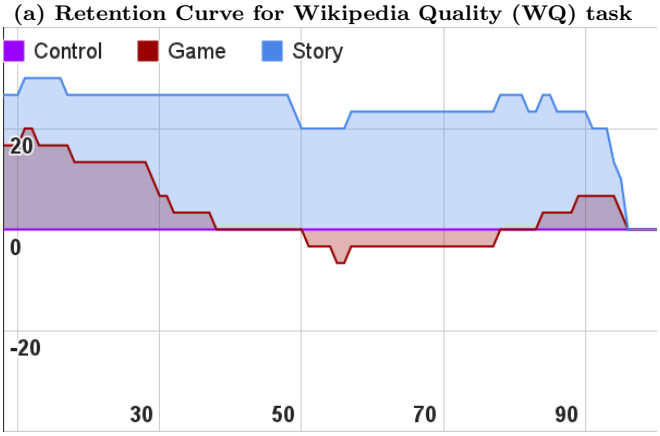
We present worker survival analysis using retention as our dependent measure. For each of the task types, we will first show retention curves and statistics, and then report on ANOVA analysis results on the log transformed *task counts* (the total number of work units completed per worker.) Finally, we show the corresponding results of a negative binomial regression, which is typically used to model repeated binary survival events.

Retention Curves

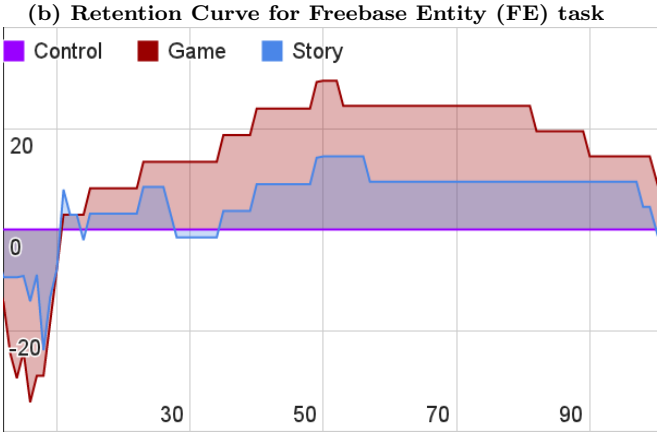
Figure 3 shows the relative change compared to the Control (No Diversions, ND) baseline for each diversion type.



(a) Retention Curve for Wikipedia Quality (WQ) task



(b) Retention Curve for Freebase Entity (FE) task



(c) Retention Curve for Image Identification (II) task

Figure 3. Retention (survival) comparison for three different task types under three different diversion types. X-axis is the number of tasks finished by the workers in a task-diversion condition (longer is better). The Control is the baseline and normalized to be flat on the x-axis. Y-axis is the percentage of delta workers retained as compared to the Control baseline (purple) for at least x tasks under the Game (red) and Story (blue) diversion types (higher is better).

The x-axis depicts the total number of tasks completed, and the y-axis shows the percentage difference of the amount of workers who *survived*, or chose to work on more tasks, after finishing a certain number of tasks (x), as compared to the Control baseline. So higher on the y-axis and further to the right on the x-axis are both better. For the three different diversion types, the Control ND is the purple baseline and normalized to be flat on the X-axis, while Game Diversion (GD) type is red, and Story Diversion (SD) is blue.

Wikipedia article Quality rating (WQ) tasks: As shown in Figure 3(a), we find that both the Game and the Story micro-diversion types influenced workers to perform more work when compared to the Control Baseline. More than 20% of workers stayed compared to the Control after finishing half of the tasks. Since workers are essentially paid the same across these different diversion conditions, this is an impressive difference.

Freebase Entity merging (FE) tasks: As shown in Figure 3(b), we see something slightly different from the WQ tasks above. While the Story micro-diversion continues to have a significant, positive impact on worker engagement, the Game condition converges to the Control baseline asymptotically. We surmise that this is due to the cognitive nature of the FE task being too different from our Game diversion, introducing switching costs by distracting workers.

Image Identification (II) tasks: As shown in Figure 3(c), in the beginning, we retained less workers in both the Game and Story diversions, suggesting micro-diversions did not help. Due to the rapid, perceptual nature of the task, it can be completed in a short amount of time with low effort. These tasks are the most profitable when someone completes them in large volume, so workers prefer to do them quickly and without interruption. Even Game condition can be a distraction, since it requires a bit of thinking and planning. Interestingly, after iterating through some instances of the tasks, micro-diversions start to keep more workers around, with Game retaining workers quite well.

Retention Statistics

We computed the means and standard deviations of the number of tasks completed. Results in Table 1 are consistent with our retention curve analysis above, and show that using diversions can improve workers' average survival rate, by a factor of 5 in the best scenario (Story Diversion with the Wikipedia Quality task).

	Wikipedia	Freebase	Image
Control	5.4 (8.6)	12.7 (26.2)	26.3 (30.3)
Game	16.7 (29.3)	16.9 (26.4)	**41.1** (44.8)
Story	**25.3** (37.0)	**35.8** (42.2)	31.85 (38.0)

Table 1. Average number of work units each worker completed for each diversion-task combination. The standard deviations are reported inside parentheses. The best results for each task type are in bold font.

ANOVA Analysis on Retention

We performed a two-way ANOVA analysis on the log transformation of the task count[2]. The results show a main effect on Task Type, $F(2, 230) = 8.400, p = .0003$, as well as a main effect on Diversion Type, but instead at the $p < .05$ level, $F(2, 230) = 4.203, p = .016$. There were no interaction effects, $F(4, 230) = 0.889$, n.s.

Post-hoc Pairwise t-test with Holm-Bonferroni correction at the $p = .01$ level of significance showed that the Image (II) task retained workers longer when compared with Wikipedia (WQ) and Freebase (FE) tasks, but WQ and FE were not significantly different from each other.

Another post-hoc Pairwise t-test with Holm-Bonferroni correction at the $p = .01$ level showed that the Story diversion type retained workers longer than Control and Game types, with Control and Game not significantly different from each other.

In short, the results here suggest that the Image (II) task generally retains workers better than Wikipedia (WQ) and Freebase (FE) tasks. Moreover, the Story type seems to be more engaging to workers than the other diversions.

Negative Binomial Regression Analysis on Retention

Using negative binomial regression, we modeled the influence of the diversions on the number of completed work units using Control with no diversions as the baseline. Results are shown in Table 2.

For the Wikipedia task (Regression 1), the results suggest that both Game and Story diversions improved retention compared to the baseline.

For the Freebase task (Regression 2), the Story diversion has retention effect with a significant coefficient, while the Game diversion does not.

For the Image task type (Regression 3), for both Game and Story conditions, while the coefficient estimates are positive, these results have not reached significance. However, the results here are still consistent with what we have observed with the Wikipedia and Freebase tasks in Regressions 1 and 2.

In general, the results suggest that micro-diversions are more effective in complex, cognitive tasks in retaining workers. Intuitively, there may be a continuum of task domains, where some tasks are so easy that the diversions are not desirable. However, as the tasks become harder, the diversions start to have an effect. Moreover, this is consistent with past literature on interruptions suggesting that context determine whether interruptions might have impact [22, 30, 34].

[2]The task count distribution can be modeled with negative binomial, therefore we applied log transformation to the task count. An alternative statistical test here is to use the non-parametric equivalent of Kruskal-Wallis test, which also resulted in similar conclusions here, with significance on Task Type, $\chi^2(2) = 16.72, p < .001$, as well as Diversion Type, $\chi^2(2) = 6.628, p = 0.036$.

	Regression 1 Wikipedia	Regression 2 Freebase	Regression 3 Image
(Intercept)	**1.6802**	**2.5416**	**3.2682**
Game	**1.0924**	0.2837	0.4478
Story	**1.5185**	**1.0354**	0.1929

Table 2. Negative binomial regression coefficient estimates using Control as the baseline and introducing either a Game or Story diversions. Significant results ($p < 0.01$) are in bold font. A positive coefficient means the condition improves worker retention.

RQ2: Worker Performance

Answer Response Time

	Wikipedia	Freebase	Image
Control	190.2 (112.5)	115.4 (111.5)	52.9 (32.5)
Game	**134.9** (98.9)	80.6 (64.1)	53.6 (43.1)
Story	145.3 (119.9)	**62.3** (86.6)	**50.0** (51.4)

Table 3. Average median time in seconds per worker. The standard deviations are reported in parentheses. The best time values per task type are in bold font.

A two-way ANOVA analysis on the log transformation of the task time yielded a main effect for Task Type at the $p < .01$ level, $F(2, 230) = 35.004, p < .001$, as well as a main effect for Diversion Type, $F(2, 230) = 6.575, p = .0017$. There were no interaction effects, $F(4, 230) = 1.292, n.s.$

Post-hoc Pairwise t-test with Holm-Bonferroni correction at the $p = .01$ level of significance showed that the Wikipedia task type took significantly more time than Freebase and Image task types, but Freebase and Image task types were not significantly different from each other.

Another post-hoc Pairwise t-test with Holm-Bonferroni correction at the $p = .01$ level showed that the Story diversion was significantly faster than the Control, with other comparisons being non-significant.

Based on these results, it seems that task contexts as well as the type of micro-diversions affect answer response time. Indeed, inspecting Table 3, the results suggest that Game and Story both improved answer response time quite a bit for the Wikipedia and Freebase task types.

Worker Accuracy

	Wikipedia	Freebase	Image
Control	66.7 (38.3)	67.8 (32.1)	79.3 (20.1)
Game	**84.2** (23.9)	**69.9** (30.2)	**81.4** (15.7)
Story	68.8 (31.0)	66.9 (33.2)	78.3 (15.3)

Table 4. Worker Accuracy in percentage (%) for each task-diversion combination. The standard deviations are reported in parentheses. The best accuracy values per task type are in bold font.

A two-way ANOVA analysis on accuracy yielded a main effect for Task Type at the $p < .01$ level, $F(2, 231) =$

64.62, $p < .001$. However, Diversion Type was not significant, $F(2, 193) = 0.550$, n.s. The interaction effect was also not significant, $F(4, 193) = 0.854$, n.s.

Post-hoc Pairwise t-test with Holm-Bonferroni correction at the .01 level of significance showed that Wikipedia and Image task types were significantly more accurate than Freebase task type, but Wikipedia and Image task types were not significantly different from each other.

The results here suggest that, while micro-diversions did not improve worker accuracy, they did not decrease worker accuracy either. Task type is mainly what determines the accuracy of the workers.

Overall, the results suggest that, with micro-diversions served at fixed frequencies, workers seems faster in completing their tasks (with the exception of the Image task) while retaining the same level of accuracy.

DISCUSSIONS

Here we return to the research questions to see what conclusions we might draw from the results of the experiment.

First, for RQ1 on retention, we saw in the experiment that micro-diversions can significantly retain workers for more work, compared to no diversions. Moreover, certain diversion types appear to retain worker better in some task domains. This is interesting, because the results suggest that we can improve worker experience by introducing micro-diversions into long, monotonous crowdsourcing tasks.

Second, for RQ2 on work performance, we saw in the experiment that workers retained their accuracy, suggesting that micro-diversions are not a significant disruptive interruption to their workflow. It is reassuring that, while micro-diversions did not actually improve worker accuracy (perhaps because workers were already as accurate as they could be), diversions were not a detriment to worker accuracy either.

More interestingly, there are evidences that workers are faster if they receive micro-diversions. In particular, the Story diversion seems to encourage workers to perform faster compared to other diversion types. If workers can maintain their accuracy while working faster, this would have implications for how we should design long, monotonous tasks on crowdsourcing platforms.

Combined with the results from RQ1, we are starting to get a picture that micro-diversions not only can retain workers for longer, but help them work faster, while retaining the same level of work quality! While we did not measure boredom directly, the results are suggestive that workers do experience boredom and that entertaining diversions might enable them to combat these effects.

Finally, for RQ3 on contextual effects, we seek to understand whether specific combinations of task and diversion type might work better than other combinations. While the ANOVA results suggest there is no interaction effects, the negative binomial regression suggests that certain combinations seem to work better than others. Theoretically, since task count is a negative binomial distribution, the negative binomial regression should yield a better model than ANOVA. From this perspective, the results of the regression analysis are worthy of further investigation.

Intuitively, we believe it makes sense that particular diversion types work better with certain task domains. The psychology of tasks suggest that humans are able to deal with fast, instinctive, perceptual tasks separately from the more deliberative, logical, cognitive tasks [24]. The Image identification tasks correspond to the faster, perceptive tasks discussed in Kahneman's work, while the Wikipedia tasks correspond to the more slower, cognitive tasks. Based on our results, more research is necessary to fully understand how these two task types interact, when either type can be the main task or the micro-diversion.

Interestingly, we did not investigate the case where the micro-diversions could also result in useful work. MTurk workers currently curate their own mix of tasks by searching over the available tasks, such that the mix remains interesting to them. Our research suggest some automation might be possible to maximize worker productivity and engagement.

CONCLUSIONS

In this paper, we proposed introducing micro-diversions to crowdsourcing workflows to improve worker experience, so as to retain good quality workers for longer. We implemented an experimental system for delivering micro-diversions interleaved with tasks to crowdsourced workers in a controlled fashion.

We explored different types of micro-diversions, including a wager-based game and a serialized webcomic story. We also examined several different task types, including a Wikipedia quality rating task, an Image identification annotation task, and a Freebase entity resolution task.

We demonstrated that introducing micro-diversions into a workflow can significantly improve worker retention as well as their answer speed, while retaining the same level of work quality. The improved worker retention and performance varied depending on the type of task and micro-diversion they received. Interestingly, in our experiment, tasks and diversions that are more closely aligned in context seemed to perform better. Moreover, the results also suggest that complex cognitive tasks might benefit from micro-diversions more than fast perceptual tasks.

As far as we know, this is the first attempt to improve crowdsourced worker engagement through timely micro-diversions. In the future, we plan to study different reward amount as well as the nature and complexity of the task versus the effectiveness of micro-diversions. We surmise that micro-diversions are more useful when they

employ different thinking systems than the main task. We also plan to apply machine learning and decision-theoretic models [15, 33] to dynamically recognize workers' need for a refresh break and schedule them automatically.

REFERENCES

1. Piotr D. Adamczyk and Brian P. Bailey. If not now, when?: The effects of interruption at different moments within task execution. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 271–278, 2004.
2. James Bo Begole and John C. Tang. Incorporating human and machine interpretation of unavailability and rhythm awareness into the design of collaborative applications. *Human-Computer Interaction*, 22(1-2):7–45, 2007.
3. Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. Soylent: A word processor with a crowd inside. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST, 2010.
4. Kurk Bollacker, Colin Evans, Praveen Paritosh, Time Sturge Tim, and Jamie Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250, 2008.
5. Chris Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, 2009.
6. Jonathan S. A. Carriere, J. Allan Cheyne, and Daniel Smilek. Everyday attention lapses and memory failures: The affective consequences of mindlessness. *Consciousness and Cognition*, 17:835–847, 2008.
7. Jonathan S. A. Carriere, J. Allan Cheyne, and Daniel Smilek. Boredom: An emotional experience distinct from apathy, anhedonia, or depression. *Social and Clinical Psychology*, 30(6):647–666, 2011.
8. Dana Chandler and Adam Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *CoRR*, abs/1210.0962, 2012.
9. Lydia B. Chilton, John J. Horton, Robert C. Miller, and Shiri SAzenkot. Task search in a human computation market. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 1–9, 2010.
10. Mihaly Csikszentmihalyi. *Creativity: Flow and the Psychology of Discovery and Invention*. HarperCollins, 1997.
11. Mihaly Csikszentmihalyi. *Beyond Boredom and Anxiety: Experiencing Flow in Work and Play*. Jossey-Bass, 2000.
12. Edward Cutrell, Mary Czerwinski, and Eric Horvitz. Notification, disruption, and memory: Effects of messaging interruptions on memory and performance. In *INTERACT*, pages 263–269, 2001.
13. Mary Czerwinski, Edward Cutrell, and Eric Horvitz. Instant messaging and interruption: Influence of task type on performance, 2000.
14. Peng Dai, Christopher H. Lin, Mausam, and Daniel S. Weld. POMDP-based control of workflows for crowdsourcing. *Artificial Intelligence*, 202:52–85, 2013.
15. Peng Dai, Mausam, and Daniel S. Weld. Decision-theoretic control of crowd-sourced workflows. In *AAAI*, 2010.
16. Peng Dai, Mausam, and Daniel S. Weld. Artificial intelligence for artificial artificial intelligence. In *AAAI*, 2011.
17. Pinar Donmez, Jaime G. Carbonell, and Jeff G. Schneider. A probabilistic framework to learn from multiple annotators with time-varying accuracy. In *SDM*, pages 826–837, 2010.
18. Robert A. Henning, Steven L Sauter, Gavriel Salvendy, and Edward F. Krieg. Microbreak length, performance, and stress in a data entry task. *Ergonomics*, 32(7):855–864, 1989.
19. John Joseph Horton. Employer expectations, peer effects and productivity: Evidence from a series of field experiments. *CoRR*, abs/1008.2437, 2010.
20. John Joseph Horton and Lydia B. Chilton. The labor economics of paid crowdsourcing. *CoRR*, 2010.
21. Eric Huang, Haoqi Zhang, David C. Parkes, Krzysztof Z. Gajos, and Yiling Chen. Toward automatic task design: A progress report. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 77–85, 2010.
22. Shamsi T. Iqbal and Brian P. Bailey. Effects of intelligent notification management on users and their tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 93–102, 2008.
23. Shamsi T. Iqbal and Eric Horvitz. Disruption and recovery of computing tasks: Field study, analysis, and directions. In *CHI*, pages 677–686, 2007.
24. Daniel Kahneman. *Thinking, Fast and Slow*. Macmillan, 2011.

25. Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 453–456, 2008.

26. Aniket Kittur, Susheel Khamkar, Paul André, and Robert Kraut. Crowdweaver: Visually managing complex crowd work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW, pages 1033–1036, 2012.

27. Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, pages 43–52, 2011.

28. Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He says, she says: Conflict and coordination in wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI, pages 453–462, 2007.

29. Shailesh Kochhar, Stefano Mazzocchi, and Praveen Paritosh. The anatomy of a large-scale human computation engine. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 10–17, 2010.

30. John G. Kreifeldt and M. E. Mccarthy. Interruption as a test of the user-computer interface. In *Proceedings of the 17th Annual Conference on Manual Control*, pages 655–667, 1981.

31. Gerald P. Krueger. Sustained work, fatigue, sleep loss and performance: A review of the issues. *Work & Stress*, 3:129–141, 1989.

32. Chris J. Lintott, Kevin Schawinski, Kate L, Daniel Thomas, M. Jordan Raddick, Robert C. Nichol, Dan Andreescu, Phil Murray, and Jan Van Den Berg. Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey," monthly notices of the royal. *Astronomical Society*, 2008.

33. Andrew Mao, Ece Kamar, and Eric Horvitz. Why stop now? predicting worker engagement in online crowdsourcing. In *HCOMP*, 2013.

34. Gloria Mark, Daniela Gudith, and Ulrich Klocke. The cost of interrupted work: More speed and stress. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 107–110, 2008.

35. Nathalie Pattyn, Xavier Neyt, David Henderickx, and Eric Soetens. Psychophysiological investigation of vigilance decrement: Boredom or cognitive fatigue? *Physiology & Behavior*, 93:369–378, 2008.

36. Layne P. Perelli. *Fatigue Stressors in Simulated Long-duration Flight: Effects on Performance, Information Processing, Subjective Fatigue, and Physiological Cost*. USAF School of Aerospace Medicine, Aerospace Medical Division (AFSC), 1980.

37. George Robertson, Eric Horvitz, Mary Czerwinski, Patrick Baudisch, Dugald R. Hutchings, Brain Meyers, Daniel Robbins, and Greg Smith. Scalable fabric: Flexible task management. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, pages 85–89, 2004.

38. Jeffrey M. Rzeszotarski and Aniket Kittur. Instrumenting the crowd: Using implicit behavioral measures to predict task performance. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST, pages 13–22, 2011.

39. Cheri Speier, Joseph S. Valacich, and Iris Vessey. The effects of task interruption and information presentation on individual decision making. In *ICIS*, pages 21–36, 1997.

40. Digger U. Verson, 2012. www.diggercomic.com.

41. Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier R. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *NIPS*, pages 2035–2043, 2009.

42. Glenn Wylie and Alan Allport. Task switching and the measurement of "switch costs". *Psychological Research*, 63:212–233, 2000.

43. Lixiu Yu and Jeffrey V. Nickerson. Cooks or cobblers?: Crowd creativity through combination. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1393–1402, 2011.