

# Inserting Micro-Breaks into Crowdsourcing Workflows

Jeffrey M. Rzeszutarski<sup>1,2</sup>, Ed Chi<sup>1</sup>, Praveen Paritosh<sup>1</sup>, Peng Dai<sup>1</sup>

<sup>1</sup> Google, Mountain View, CA 94043, USA

<sup>2</sup> Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

## Abstract

Participants in human computation workflows may become fatigued or get bored over long, interminable working hours. This leads to a slump of motivation and morale, which in the long run causes reductions in both productivity and work quality. In this paper we propose an initial investigation into possible ways to alleviate worker fatigue and boredom by employing micro-breaks that provide timely relax to workers during long sequences of tasks. We experimentally test micro-breaks on Amazon's Mechanical Turk, showing that micro-breaks can significantly improve worker retention rate as task batches reach hours in length, and appear to increase overall worker engagement and commitment to their work.

## Introduction

In crowdsourcing task markets such as Amazon's Mechanical Turk or oDesk, hundreds of thousands of workers take on tasks that are difficult for a computer to solve. Workers on such markets can often perform a large amount of tasks in one huge batch, which in many ways is efficient and beneficial due to worker expertise growth.

On the other hand, many workers choose to work long hours, perhaps for reasons such as economics, competitive pressure from peers, or bad habits. We intuitively know that high task loads and long working hours can introduce negative side effects such as boredom and fatigue. For example, imagine an easy task that has workers tag images. A worker can easily tag the images or even do a few more without trouble. Suppose a worker has been tagging continuously for four hours. She has become extremely bored, and her reduced attention may not even meet the demands of such a simple task. Likewise, another worker may take shortcuts by cutting and pasting sets of tags. These working strategies are potential coping mechanisms for reduced cognitive resources and boredom.

While we now know how to mitigate quality control issues in crowdsourcing systems using CAPTCHA questions, gold standards, and majority voting (Bernstein et al. 2010; Callison-Burch 2009), it is much less clear how to deal with other performance issues such as reduced cognitive abilities brought on by fatigue and boredom. We know that fatigue, both physical and cognitive, can affect workers doing large batches of tasks, not only risking strain injuries and reduced well-being, but also creating lower quality, unreliable data as a result (Krueger 1989). Even worse, many quality control measures for crowdsourcing tasks are guided by consensus, which may be forfeited by unreliable user performance.

We propose a method that operates within a human computation system to help mitigate fatigue and boredom through breaks that are similar in size and scope to the other human computation tasks. Imagine a worker has been rating the quality of search results for 15 minutes. All of the sudden she gets a micro-break, and instead of evaluating another page she sees encouraging text and a leaderboard showing how well she is doing. Or, instead, she is awarded a funny video for 30 seconds. Such interruptions do indeed incur a cost in context switching, as the worker must return to her task afterwards (Wylie and Allport 2000). Yet, the interrupting refreshes the worker.

## Study

We performed an initial experimental investigation that explores whether implementing micro-breaks in crowdsourcing markets (Amazon's Mechanical Turk) improves worker engagement. We implemented a system for delivering micro-breaks in a controlled fashion using a break scheduler built with Google AppEngine.

Research shows that breaks can ameliorate boredom, but that their effects vary. Breaks may have more impact if they are similar to the task, and generally, the benefits for a break roughly match the length of the break (Henning et al. 1989). But even small three- to thirty-second breaks can have a benefit. We developed two different breaks based

	Wikipedia	Merging	Images
<i>intercept</i>	1.6802 p<0.001***	2.5416 p<0.001***	3.8622 p<0.001***
<i>Game break</i>	1.0924 p=0.002**	0.2837 p=0.454	-0.7662 p=0.042*
<i>Comic break</i>	1.5185 p<0.001***	1.0354 p=0.006**	0.4013 p=0.285

Table 1: Negative binomial regression coefficient estimates for game and comic conditions.

on these findings with a goal of maximizing potential effects and exploring the design space of micro-breaks.

In our first break condition we give workers a game where they can choose to risk part of their current earnings for a fair chance of more payout. We adjust the odds so that the expected extra earnings are zero, yet because participants may win from time to time, the gambler’s fallacy may provide a degree of extrinsic motivation. In the second break condition we give workers an eye-catching comic to read. We chose these breaks because one is primarily motivation and judgment-based while the other is non-directed and cognitive. This may cause different interaction effects with differing tasks.

Our tasks include a Wikipedia article evaluation task that requires reading comprehension, a knowledge base entity merging task that requires studying and making a judgment, and an image subject identification task that needs only simple, binary judgments in rapid succession. We populated queues with at least an hour of work so as to induce boredom and fatigue. We also introduced gold standard data so that we could measure work quality.

We recruited 30 unique workers per condition (3 tasks X 2 breaks + control/no break) to stay and do as many copies of the different tasks as we had. This gives us 3x3x30=270 unique worker submissions. We adjusted the payment rate and had breaks appear within tasks so that the compensation was equal between conditions. We examined the retention rate of the workers to measure break effectiveness. If more workers stayed to do more tasks in a condition, we suggest that it was more engaging.

## Results

For the Wikipedia rating task, we observed that both the game and comic micro-breaks influenced workers to perform more work comparing to the control baseline. More than 20% of workers were still around compared to the control case after doing half of the tasks, yet without increasing their total pay. Negative binomial regression (used because of the exponential distributions) suggests that the breaks had a significant effect (Table 1). In the entity merging tasks we see something slightly different. While the comic micro-break still has an impact, the game condition converges to the control baseline. This is confirmed by a negative binomial regression. We surmise that this is due to the cognitive nature of the task being too

different from the game, introducing switching costs and making the break less effective.

The image identification task is markedly different. Micro-breaks seem to have a negative effect. Instead of keeping people around, they seem to be pushing people away. This is possibly due to the rapid, perceptual nature of the task, as many tasks can be completed in a short amount of time with low effort. Because even the game requires some effort, it might only serve as a distraction. Our analysis shows that the negative effect on retention is in fact significant in the game condition. Overall, we notice slight time and accuracy improvements with breaks, though no significant changes.

## Discussion

This work makes several contributions. We propose introducing micro-breaks to crowdsourcing workflows to overcome boredom and fatigue. We describe the implementation of such a system, and provide example breaks. We conduct experiments demonstrating that introducing micro-breaks into a workflow can make a difference in worker retention rates and slightly improves accuracy and work time. We find the difference in retention varies depending upon the type of task workers were doing, and the kind of micro-break they received. Tasks and breaks that were more closely aligned in content seemed to perform better, and the results suggest that increasing the complexity of a task improved the effectiveness of the micro-breaks. In the future, we plan to study more thoroughly reward amount and the nature of the task versus the effectiveness of micro-breaks. We suppose that micro-breaks are more useful when it is relatively less competitive with the main task and cognitively challenging. We plan to apply machine learning and decision-theoretic models (Dai et al. 2010) to optimally schedule breaks and suggest break types.

## References

- Bernstein, M.S., Little, G., Miller, R.C., et al. 2010. Soylent: a word processor with a crowd inside. In Proc. UIST ’10, 313–322.
- Callison-Burch, C. Fast, cheap, and creative: Evaluating translation quality using Amazon’s Mechanical Turk. 2009. In Proc. Conference on Empirical Methods in NLP, 286–295.
- Dai, P., Mausam, and Weld, S. 2010. Decision-theoretic control of crowd-sourced workflows. In Proc. AIII ’10.
- Henning, R., Sauter, S., Salvendy, G., and Kreig, E. 1989. Microbreak length, performance, and stress in a data entry task. *Ergonomics* 32, 7, 855–864.
- Krueger, G. 1989. Sustained work, fatigue, sleep loss and performance: A review of the issues. *Work & Stress* 3,2,129–141.
- Wylie, G. and Allport, A. 2000. Task switching and the measurement of “switch costs”. *Psychological research* 63, 3-4.