Using Microtask Continuity to Improve Crowdsourcing

Walter S. Lasecki[1], Adam Marcus[2], Jeffrey M. Rzeszotarski[3], and Jeffrey P. Bigham[3]

Computer Science, ROC HCI[1]
University of Rochester
wlasecki@cs.rochester.edu

Locu Inc.[2]
marcua@cs.cmu.edu

HCII[3]
Carnegie Mellon University
{jeffrz,jbigham}@cmu.edu

ABSTRACT

A rich body of cognitive science literature suggests that workers who focus on a single task in a large workflow leverage task specialization to improve the overall performance of the workflow, such as in an assembly line. However, crowdsourcing workflows often ignore worker growth over time, instead treating them as homogeneous computational units that can effortlessly move between small *microtasks* of different types. In this paper, we validate that workers often mix different task types via a survey, and then study the effects of such task type mixing. We collect empirical evidence from 338 crowd workers that suggests task interruptions significantly decrease worker performance. Specifically, we show that temporal interruptions, where there is a large delay between two tasks, can cause up to a 102% slowdown in task completion time, and contextual interruptions, where workers are asked to perform different tasks in sequence, can slow down completion time by 57%. Our results demonstrate the importance of considering continuity in workflow design for both individual worker efficiency and overall throughput.

Author Keywords

Crowdsourcing, human computation, workflows, continuity, interruptions, efficiency

ACM Classification Keywords

H.4.2 Information Interfaces & Presentation: User Interfaces

INTRODUCTION

Workflow design has been an important topic since before the advent of the assembly line because of its financial and societal impact. Much of the work in this area has supported letting workers specialize in specific tasks, and avoiding interruptions to the workflow, as a means of improving efficiency and overall productivity [24]. In contrast, crowdsourcing, which has quickly risen in popularity for its ability to solve tasks that artificial intelligence cannot yet solve, largely fails to recognize the impact of prior work indicating that reducing shifts in focus or context is more efficient. When a large workflow is broken down into small *microtasks*, it often encourages workers to work on different types of tasks by introducing near-equivalent overhead for continuing to do the same tasks. Also, because crowd workers can easily move between tasks from different requesters, it is difficult to encourage specialization. Despite these potential issues, prior work in crowdsourcing has focused on efficiency mainly via algorithmic optimizations of the cost of work.

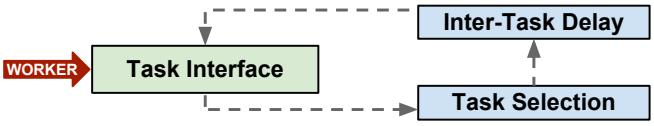

Figure 1. Our experiments measure how interleaving tasks and adding inter-task delays impact the efficiency of workers. To test this, we simulate interruptions to both the new task selection (contextual) and the between-task interval (temporal).

In this paper, we demonstrate the benefit of taking into account prior work in cognitive science and psychology, which points to continuity (lack of interruption) being beneficial for productivity, in crowdsourced workflow design. We identify and quantify two types of interruptions that are harmful to worker efficiency: 1) *contextual interruptions*, in which workers swap between tasks of different types, and 2) *temporal interruptions*, in which workers must wait between submitting one task and receiving the next one. While contextual interruptions can come as a result of workflows in which workers are asked to switch between task types, they can also be caused by workers who diversify their task selection. Temporal interruptions, on the other hand, are generally in the hands of workflow designers. In a survey of 100 Mechanical Turk workers, we find that contextual interruptions are common as a result of current design patterns in which workers often switch between unrelated tasks. As a consequence of this behavior, it is critical to quantify the effect of various types of interruptions to ensure that workflow designers do not inadvertently create more interruptions than necessary.

With this basis in mind, we designed an experiment to test the impact of temporal and contextual interruptions on workers, and collected empirical evidence from 338 workers that both types of task interruptions can have significant effects on task performance, and ultimately the cost of a workflow. We show that both temporal and contextual interruptions are harmful to workflow latency. These findings suggest that continuity in both receiving new tasks and the context of new tasks is an important factor in the productivity of crowdsourced workflow design. Thus, by reducing interruptions, workers can

earn higher hourly wages for their work, and task owners can receive results more efficiently. We then discuss a set of steps that task designers can take to minimize the interruptions that workers are forced to deal with.

The contributions of this paper are as follows:

- We discuss relevant cognitive science and psychology literature to highlight prior findings that longer, less-interrupted workflows are beneficial to workers.
- We present the results of a survey that shows workers frequently switch between different types of tasks.
- We quantify the effect of interruptions, and find that temporal interruptions can slow down task completion time by 102%, while contextual interruptions can slow down task completion time by 57%, suggesting continuity might be an effective means of increasing worker efficiency.

RELATED WORK

Our work focuses on the intersection of cognitive science, which has explored the benefits of continuous (uninterrupted) workflows on an individual level, and crowdsourcing, which has focused on a highly divided microtask model to more effectively leverage open calls to non-expert workers.

Cognitive Factors in Workflow Design

The literature on task flows in psychology and cognitive science has primarily focused on the costs and benefits associated with *interruptions* [20]. Interruptions have been found to decrease the performance of a worker shortly after the interruption [12, 19]. The costs of interruptions vary depending on the nature of the work being done, the type of interruption, and the environment of the worker. Interruptions that are more closely aligned with the kind of work that a worker is already doing are likely to be more disruptive regardless of their length. Even if people are given a chance to rehearse their prior task, they struggle to resume it [10].

Moreover, workers perform poorly on the interrupting task as well [4]. Cutrell, Czerwinski, and Horvitz explored the ways instant messaging interruptions can affect a list search task [5, 6, 7]. They demonstrated that certain interaction events such as typing, evaluating a list of search criteria, or using menus were especially harmful to interrupt. More generally, interruptions have costs based on the cognitive load of the task at hand [1]. While this does not directly match the situation of microtask workers, it sheds light on the potential cognitive costs of interrupting an active workflow, and suggests that interruptions in crowdsourcing workflows are worth exploring.

When people are disrupted, they may lose some critical bits of information and have to repeat a part of their task. For example, people who dial a phone while driving have to take moments to reexamine the road and adjust their course during interruptions [11]. In the case of crowdsourcing, one might imagine the sorts of repeated sense-making that workers have to do to interleave tasks (e.g., re-read directions, re-learn how to complete a task, or re-develop some domain expertise). Many workers work on and switch between tasks. Multitasking is known to have many of the costs of interruption [23]. This suggests that crowdsourcing can be improved by encouraging workers to focus on one type of task for longer spans.

Crowdsourcing

Since crowd workers are fleeting and generally unknown to the requester, typical designs ensure that no prior knowledge on a task is needed so that each one can be completed by a different, untrained worker. Accordingly, most work in crowdsourcing has focused on decomposable problems such as writing and editing [2], and image description and interpretation [3, 26], among others. Existing workflows focus on obtaining quality results from workers, and generally introduce redundancy and verification across multiple workers (such as in answer agreement [26] or the find-fix-verify pattern [2]).

While this approach maximizes the flexibility of the workforce itself by not requiring prior experience or long work terms, it disregards benefits like worker experience and memory. Because subsequent tasks are unrelated, the contextual disruptions cause a loss in specialization. Prior work has shown that despite often completing dozens of tasks per hour, workers remember task-specific details [18]. This means that discrete tasks often fail to leverage the experience workers gain over the course of completing multiple tasks [8]. Additionally, discretizing microtasks can cause a temporal disruption between the submission of one task and loading a subsequent task, which might cause workers to lose interest, move to another task completely, or earn less money for their time.

Continuous Crowdsourcing

For tasks requiring ongoing input from workers, Legion [14] introduced continuous real-time crowdsourcing to engage workers with continuous tasks for longer periods of time. Continuous workflows reduce contextual interruptions from accepting new micro-units of work, and help workers build up relevant experience. For example, transcribing a continuous stream of audio might allow a worker to use their knowledge that a previous sentence mentioned "President Obama" to transcribe "The President" rather than "The precedent" if they have trouble hearing a portion of an audio clip [13]. Similar benefit can be seen in activity recognition domains as well, where workers can better infer what action is being performed in a video if they are able to see the context surrounding it, instead of just a short clip of the action itself [15].

Continuity also provides the chance to improve the end-user experience too, by leveraging the context workers maintain. For example, VizWiz [3] asks crowd workers to answer visual questions about blind users' surroundings, but often struggles to answer questions that involve prior context, such as follow-up questions, because workers are only engaged for single tasks, and routed randomly to new questions when they take a new one. Chorus:View [16] improved on VizWiz for these types of questions by allowing workers to view streaming video from the end-user's mobile device, and then engage in an ongoing conversation about the content of the video using Chorus [17], a conversational interface powered by the crowd. This let workers maintain context and use that knowledge to make it easier for the user to find the information they were looking for.

WORKER SURVEY

To better understand the worker-centric factors that affect crowdsourcing workflows, we conducted a survey of 100 Mechanical Turk workers. 55 respondents identified as male and 44 as female. 55 were aged 18-29, 34 aged 30-39, and 11 aged 40+. 85 respondents had a university degree. 58 workers identified Mechanical Turk as a major source of income.

Worker Habits

Workers spent an average of at least four hours a day ($M = 4.47$, $SD = 2.76$) working on Mechanical Turk tasks. There was a significant difference between the working hours of those doing tasks for pocket money, a part of their living, and as a job ($F(2, 97) = 4.99$, $p < 0.01$). They worked 3.5, 5, and 5.41 hours on average, respectively. 64 respondents took breaks, although we could find no relationship between why workers work and whether they take breaks at all. Those who did take breaks said they worked for an average of 1.24 hours ($SD = 2.68$) and then took a break averaging 16.6 minutes ($SD = 16.9$). Of the workers who did not take breaks, 64% of them mentioned prioritizing money as their reason, and breaks hurting their bottom line. 11% cited issues with "flow" when transitioning between tasks, 14% cited being satisfied with natural breaks as new HITs were loaded, and 11% cited time or task demands preventing them from breaking.

In addition to temporal interruptions, we also found that 77% of workers "frequently" switched tasks. They cited boredom, insufficient pay, difficulty, and too high attention demands as potential reasons for skipping to a new kind of HIT. Many of these reasons were influenced by the design of the tasks itself.

Potential for Specialization

Our study also suggests that workers remember the kinds of HITs they complete, even long after they did them, suggesting that task specialization in a microtask marketplace is possible, even without forcing workers to exclusively take a single type of task. Our results showed that 78% were able to write about the HIT they did right before taking the survey, and 74% could recall a HIT they really enjoyed. Many workers described HITs they completed days, weeks, or months ago. This agrees with prior work showing that workers are able to retain and apply task information to improve their performance in future tasks [18]. Workers cited novelty, ease, speed, and repeatability as traits common to preferred tasks.

EXPERIMENTS

Our survey showed that workers are subject to interrupted workflows in the current model. In our experiments, we explored two common types of potential interruptions: *temporal*, where a delay is added between tasks of the same type, and *contextual*, where different tasks are interleaved with tasks of the same type. Our goal was to determine if these types of interruptions have a measurable effect on workers' task performance.

Experimental Setup

To measure the effect of interruptions on workflows, our task asked workers to identify places on a map. Each map was larger than the user's viewport, forcing them to explore to find the target. We generated one random street map per worker so that over time they could remember places and geography. While we randomly placed the landmarks, all maps had the same ones, such as a school, a place of worship, and a general store. These landmarks were distributed evenly in random locations on each user's map. An example landmark map can be seen in Figure 2. For clarity and to avoid cultural bias, we used textual labels for landmarks rather than graphical icons.

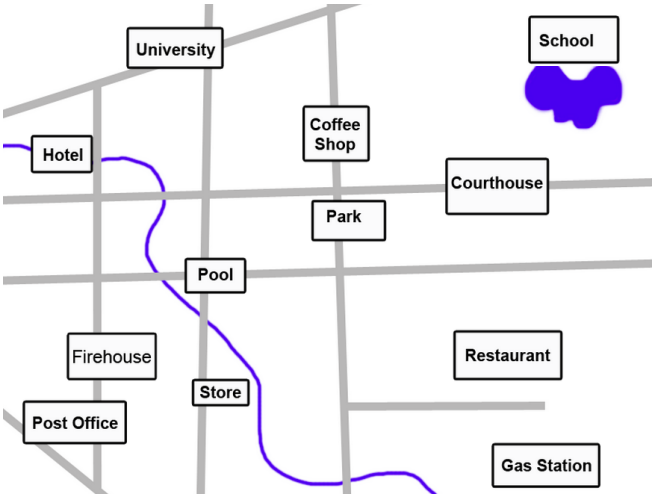

Figure 2. One of the map configurations generated for our trials. Workers were only able to see roughly $1/5^{th}$ of the map at a time through their viewport, meaning they have to scroll to find their target. Since each worker saw the same map for each of the primary search tasks, it was possible to learn the locations of the buildings over time.

In all of our experiments, we varied the amount or type of time-based or context-based disruptions. Since our tasks adhere well to the classic microtask model, in which no prior expertise or context is required, most workers were eventually able to get the correct answer. Because of this, our measures focused on how long it took workers to achieve the task, rather than the degree to which they were correct.

Different disruptions may cause workers to take different amounts of time to complete a task. We collected responses from 338 unique Mechanical Turk workers, and measured how long it takes them to complete tasks in different conditions and report what this means for throughput in terms of work lost. Task latency is our key measure because it captures how mentally prepared workers were upon seeing a new task, and how their memory of the task improved performance.

Control Task

In the control task, workers were first prompted to find and click on a particular landmark. Subsequent tasks utilized the same map, but asked workers to find different landmarks. For example, a worker might be asked to click on the general store, then after correctly clicking on the general store, asked to click on the school. We expect that as workers are asked to use their map more and more, they will become more familiar with the layout of the map, and thus able to find targets more quickly. To avoid potential biases, each worker assigned to the condition sees a different, randomly generated map.

Temporal Interruptions

Traditional microtask interfaces often force workers to pause between tasks as a new task loads. To understand the effect of this delay, we modified the amount of time a worker had to wait between successfully completing one task and being given their next task, so that it is not instantaneous as in the control task (C, $N = 57$). We used two delay lengths: 10 seconds (C_{short}, $N = 71$) and 30 seconds (C_{long}, $N = 67$).

We did not find a significant difference between the total times workers spent finding landmarks in C and in C_{short} ($t(70) = 1.19$, $p = 0.24$). This leads us to believe that short, 10 second breaks do not have a detectable effect given the sensitivity of our measures. However, longer, 30 second breaks do have a significant effect ($t(66) = 3.40$, $p < 0.01$). With longer delays, workers may become bored, might turn to distractions in their environment, or may try another quick task while they are waiting. Any of these possibilities have a disruptive effect on their working knowledge of the landmark map, and this bears out in their completion time (Figure 3).

Contextual Interruption

Microtask platforms like Mechanical Turk allow workers to multiplex tasks, loading multiple HITs in multiple browser windows or tabs. Additionally, task designers might provide contextually related or unrelated tasks to workers performing a HIT of a particular type. We measure the effects of these contextual interruptions by asking workers to complete a different task between each new location-finding task.

In the first condition (C_{map}, $N = 84$), workers who successfully identified a landmark were prompted to identify a new landmark situated on a *different* map than before. In the second condition (C_{image}, $N = 59$), workers who successfully identified a landmark were prompted to complete the unrelated task of image labeling (analogous to a worker choosing a different task between two similar ones). After providing a short description of an image, workers were prompted to find another landmark on the same map as before.

For the image description distractor task, we did not find a significant difference between C and C_{image} ($t(58) = 0.31$, $p = 0.75$). As we expect from prior work, this is likely because the image description task was short and relatively disjoint from the map identification task, it does not appear to interfere with participants' performance in locating landmarks. However, the C_{new} condition, where workers were interrupted with a new map, showed a significant effect ($t(83) = 4.39$, $p < 0.001$). Since this interruption was similar to the main task, it was more likely to interfere with the workers' knowledge.

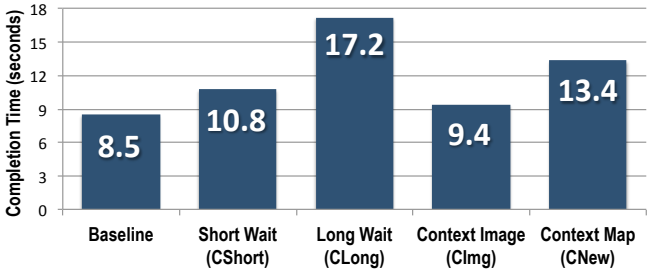

Figure 3. Results of our experiments. There is a significant delay incurred when a worker is asked to wait for a long period of time, or to interleave different instances of a similar problem.

DISCUSSION

Our results demonstrate that there exists a quantifiable difference in performance when sequential tasks are interrupted. While the individual effects we measured initially look modest, their cumulative effect in a microtask setting can be quite large. For example, it took 2.02 times as long on average for a worker to complete each landmark identification task when interrupted with a similar map (C_{map}) because the participants' working knowledge and context were potentially disrupted. This means that over an 8 hour work day, such a worker would only complete what a worker using our baseline setup (condition C) would accomplish in 3:57 hours (that is, they would only be 49.5% as efficient).

Likewise, in our delay condition (C_{long}), even if the 30 second delay itself is considered unavoidable in the workflow, it has a potential side effect, slowing the work conducted during active periods by a factor of 1.57. This means that workers would only be 63.7% as efficient as a worker in the baseline condition (ignoring all of the time spent waiting), potentially reducing their effective wage significantly. While this extrapolation is speculative, it demonstrates the impact that small changes could potentially have on workers.

FUTURE WORK

In addition to increasing task efficiency, task continuity may hold other benefits to both requesters and workers. For example, studies of expert performers suggest that an active feedback loop during practice can greatly improve a person's performance [9]. Continuous crowdsourcing systems such as Legion [14] or Legion:Scribe [13] Additionally, while interruptions can be detrimental, appropriately-timed breaks after long periods of work can provide some benefit [21, 22]. Speier, Valacich, and Vessey found that during tasks that don't tax users' mental abilities, such as simple data entry, such breaks can help workers to pay more attention [25]. These beneficial interruptions might be incorporated into continuous workflows, perhaps through status screens and small breaks. In general, leveraging prior work in cognitive science and other fields allows us to re-evaluate how the nature of crowdsourcing impacts the performance of workers, and take steps toward more efficient processes.

CONCLUSIONS

In this paper, we have discussed problems associated with lack of task continuity in existing crowdsourcing settings, and showed that workers are subject to task interruption via a survey of 100 crowd workers. We then experimentally showed the effects of temporal and contextual interruptions on workers. Our findings from 338 workers indicate that there is a significant detrimental effect on worker performance when these types of errors are introduced. We concluded with a discussion of how work in cognitive science and similar fields will allow us to improve the efficiency of the crowd in the future by creating workflows that take human nature into account.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under awards #IIS-1149709 and #IIS-1218209, Google, and two Microsoft Research Ph.D. Fellowships.

REFERENCES

1. Adamczyk, P. D., and Bailey, B. P. 2004. If not now, when?: the effects of interruption at different moments within task execution. In *CHI 2004*.
2. Bernstein, M. S.; Little, G.; Miller, R. C.; Hartmann, B.; Ackerman, M. S.; Karger, D. R.; Crowell, D.; and Panovich, K. 2010. Soylent: a word processor with a crowd inside. In *UIST 2010*, 313–322.
3. Bigham, J. P.; Jayant, C.; Ji, H.; Little, G.; Miller, A.; Miller, R. C.; Miller, R.; Tatarowicz, A.; White, B.; White, S.; and Yeh, T. 2010. Vizwiz: nearly real-time answers to visual questions. In *UIST 2010*, 333–342.
4. Cabon, P., Coblentz, A., and Mollard, R. 1990. Interruption of a monotonous activity with complex tasks: effects of individual differences. In *Human Factors Soc. Meeting 1990*.
5. Cutrell, E., Czerwinski, M. and Horvitz, E. 2001. Notification, disruption, and memory: Effects of messaging interruptions on memory and performance. In *INTERACT 01*, 263-269.
6. Czerwinski, M., Cutrell, E., and Horvitz, E. 2000. Instant messaging and interruption: Influence of task type on performance. In *OZCHI 2000*.
7. Czerwinski, M., Cutrell, E., and Horvitz, E. 2000. Instant messaging: Effects of relevance and time. In *People and computers XIV: Proceedings of HCI 2000*, Vol. 2, British Computer Society, 7176.
8. Dow, S., Kulkarni, A., Klemmer, S., Hartmann, B. 2012. Shepherding the Crowd Yields Better Work. In *CSCW 2012*.
9. Ericsson, K. A. and Krampe, R. T. and Tesch-Romer, C. 1993. The role of deliberate practice in the acquisition of expert performance. In *Psych. Review*, 100.3, 363.
10. Gillie, T. Broadbent, D. 1989. What Makes Interruptions Disruptive? A Study of Length, Similarity, and Complexity. *Psych Research*, 50, 243-50.
11. Jansset, C. P., and Brumby, D. P. 2010. Strategic Adaptation to Performance Objectives in a DualTask Setting. *Cog. Sci.*, 34.8, 1548-60.
12. Kreifeldt, J. G., and McCarthy, M. E. 1981. Interruption as a test of the user-computer interface. In *Manual Control*, JPL Publication, 8195.
13. Lasecki, W. S.; Miller, C. D.; Sadilek, A.; Abumoussa, A.; Borrello, D.; Kushalnagar, R.; and Bigham, J. P. 2012a. Real-time captioning by groups of non-experts. In *UIST 2012*.
14. Lasecki, W. S.; Murray, K.; White, S.; Miller, R. C.; and Bigham, J. P. 2011. Real-time crowd control of existing interfaces. *UIST 2011*, 23–32.
15. Lasecki, W. S.; Song, Y. C.; Kautz, H.; and Bigham, J. P. Real-Time Crowd Labeling for Deployable Activity Recognition. In *CSCW 2013*.
16. Lasecki, W. S., Thiha, P., Zhong, Y., Brady, E., Bigham, J. P. Answering Visual Questions with Conversational Crowd Assistants. In *ASSETS 2013*. To Appear.
17. Lasecki, W. S., Wesley, R., Nichols, J., Kulkarni, A., Allen, J. F., and Bigham, J. P. Chorus: A Crowd-Powered Conversational Assistant. In *UIST 2013*. To Appear.
18. Lasecki, W. S.; White, S.; Murray, K.; and Bigham, J. 2012b. Crowd memory: Learning in the collective. In *Collective Intelligence 2012*.
19. McFarlane, D.C. and Latorella, K.A. 2002. The Scope and Importance of Human Interruption in HCI design. In *HCI*, 17, 1-62.
20. O'Connail, B. and Frohlich, D. 1995. Timespace in the workplace: Dealing with interruptions. In *CHI 1995*, 262-263.
21. Pattyn, N., Neyt, X., Henderickx, D., and Soetens, E. 2008. Psychophysiological investigation of vigilance decrement: boredom or cognitive fatigue?. In *Physiology and Behavior* 93, 1-2, 369-78.
22. Rzeszotarski, J.M., Chi, E., Paritosh, P., Dai, P. 2013. Inserting Micro-Breaks into Crowdsourcing Workflows. In *HCOMP 2013*, .
23. Salvucci, D.D., and Bogunovich, P. 2010. Multitasking and monotasking: the effects of mental workload on deferred task interruptions. In *CHI 2004*, 85-88.
24. Smith, A. 2004. The Wealth of Nations. *Digireads*. http://books.google.com/books?id=rBiqT86BGQEC
25. Speier, C., Valacich, J. S., and Vessey, I. 1997. The effects of task interruption and information presentation on individual decision making. In *InfoSys*, 21–36.
26. von Ahn, L., and Dabbish, L. 2004. Labeling images with a computer game. In *CHI 2004*, 319–326.