Diminished Control in Crowdsourcing: An Investigation of Crowdworker Multitasking Behavior

SANDY J. J. GOULD, University College London
ANNA L. COX, University College London
DUNCAN P. BRUMBY, University College London

Obtaining high-quality data from crowds can be difficult if contributors do not give tasks sufficient attention. Attention checks are often used to mitigate this problem, but, because the roots of inattention are poorly understood, checks often compel attentive contributors to complete unnecessary work. We investigated a potential source of inattentiveness during crowdwork: multitasking. We found that workers switched to other tasks every five minutes, on average. There were indications that increasing switch frequency negatively affected performance. To address this, we tested an intervention that encouraged workers to stay focused on our task after multitasking was detected. We found that our intervention reduced the frequency of task-switching. It also improves on existing attention checks because it does not place additional demands on workers who are already focused. Our approach shows that crowds can help to overcome some of the limitations of laboratory studies by affording access to naturalistic multitasking behavior.

Categories and Subject Descriptors: **·Human-centered computing → Empirical studies in HCI ·Human-centered computing → HCI theory concepts and models ·Information systems → Crowdsourcing**

Additional Key Words and Phrases: Interruptions; multitasking; cuing; crowdsourcing; online experimentation; methodology; human performance; data-entry; transcription

1 INTRODUCTION

Online crowdsourcing platforms have opened-up new ways of getting work done and carrying out research. Crowdsourcing has a number of benefits, such as fast response times [Kittur et al. 2008; Bernstein et al. 2011] and large sample sizes [e.g., Chandler et al. 2014]. Being a relatively new way of working, requesters (who set tasks and pay for work) have also faced challenges with crowdsourcing, one of which has been worker attentiveness.

Online crowdworkers are largely unsupervised. This can encourage satisficing: behavior where workers optimize for return on effort rather than quality [Gould, Cox and Brumby 2015]. To reduce inattentiveness, researchers have tested *attention checks*. These often take the form of extra questions added to tasks that help identify workers who aren't paying attention [e.g., Kittur et al. 2008; Walther et al. 2011; Paolacci et al. 2010; Peer et al. 2014].

Attention checks can help identify poor performers, but they often fail to address the varied behavioral roots of inattentiveness. In this paper we focus on one possible cause of drifting attention: multitasking. What role does multitasking play in crowdworking settings, where people often have competing tasks and work without direct oversight? After all, multitasking is a common cause of inattentiveness in traditional workplace settings [e.g., González and Mark 2004; Grundgeiger et al. 2010]. There is little evidence about the prevalence of multitasking during crowdwork, its effects on performance, or whether crowdworkers' multitasking behavior can be modulated. An empirically-informed understanding of these factors is a step towards a better understanding of crowdworker productivity.

In this paper we present an investigation of multitasking behavior during crowdwork. We find that multitasking is at least as common in crowdworking settings as it is in traditional working environments. Through self-report we identify some of the other activities that workers commonly switch to. We find evidence that multitasking might negatively affect the performance of crowdworkers. Finally, we introduce an efficient intervention that encourages workers to resist the temptation to switch to other tasks. This intervention, unlike many attention checks, is focused on

distracted workers and does not unnecessarily consume the time of workers who are focused on the task that they are working on.

2 RELATED WORK

2.1 Inattentiveness in crowdsourcing environments

One of the challenges of obtaining high-quality data from crowds is that workers are remote and unsupervised. This is potentially problematic because it means that workers may 'satisfice' – clicking their way through tasks without paying much attention. The remote setting of crowdwork also means there is the potential for digital and non-digital distractors that can break workers' focus, further compromising quality.

The potential for inattentiveness, whether due to satisficing or distraction, is an issue for both commercial requesters who are looking for tasks to be completed as efficiently as possible, and for researchers who are increasingly using crowds for both experimental [e.g., Heer and Bostock 2010; Kittur et al. 2008; Komarov et al. 2013; Paolacci et al. 2010; Dandurand et al. 2008; Germine et al. 2012; Gould et al. 2016] and non-experimental research [e.g., Behrend et al. 2011; Buhrmester et al. 2011].

How best to handle diminished control and mitigate confounding factors has been an important challenge for the crowdsourcing community. Attempts have been made to mitigate the problem of diminished experimental control. These approaches can be characterized as remedial or preventative. Remedial approaches involve detecting egregious deviations from normal performance and removing them from a sample [e.g., Downs et al. 2010; Rzeszotarski and Kittur 2012; Rzeszotarski and Kittur 2011; Snow et al. 2008]. Preventative approaches seek to intervene, curtailing uncontrolled influences on performance as they occur. This can involve developing interventions to increase the level of control [e.g., Kapelner and Chandler 2010; Kittur et al. 2008] or inducing participants to conform to a study's requirements [e.g., Scekic et al. 2013].

One class of preventative interventions that have been utilized to improve participant performance is that of attention checks. These checks are designed to ensure that participants are completing a task as requested. They have been used most often in crowdsourced survey studies [e.g., Paolacci et al. 2010; Downs et al. 2010]. In these studies they take the form of questions that catch-out participants who are clicking through questions at random with little thought or effort given to their responses. For example, an attention check questionnaire item might be: *"This question checks whether you are skipping questions. Select the middle option."* [Walther et al. 2011]. Trials of these attention checks have demonstrated that they are generally effective in improving the quality of questionnaire responses [e.g., Goodman et al. 2013; Peer et al. 2014].

Attention checks can also take subtler forms. Kittur, Chi and Suh [2008] asked Amazon Mechanical Turk (AMT) workers to rate Wikipedia articles on their accuracy and readability. They found that the ratings of AMT workers correlated poorly with those of expert raters. Kittur and colleagues determined that these differences in ratings were caused by some workers entering responses as quickly as they could with little care for their accuracy. The difficulty in improving the quality of responses was that the ratings were necessarily subjective – there is no clear ground truth. How can the response of someone not paying attention be distinguished from the response of someone who has given a task due consideration but has arrived at a 'wrong' answer? Kittur et al.'s solution was to add questions to the task that had objectively correct answers. Making correct responses to these answers required participants to find information that was also useful when forming ratings. Workers, knowing that nonsense answers could now easily be spotted, were much more attentive.

Kittur et al.'s [2008] solution was successful but this success came at a cost – participants took three times longer to complete the task once the extra questions were

added. Some of the additional time came from participants who, instead of rushing, completed the task properly. Much of the extra time, though, would have come from focused workers doing redundant work. In this way Kittur et al.'s intervention is similar to other attention checks; both attentive and inattentive participants are required to complete the same attention check tasks. Attention checks that are instead responsive to worker behavior might provide an opportunity to improve the focus of inattentive participants, without burdening attentive participants with unnecessary work.

The use of attention checks also points to a fundamental issue: for better or worse, requesters have little control over remote workers and their habits and routines while they are working. Is it really possible to exert control in online settings? And what are the sources of variation that confound results? Both of these questions are largely unexplored in investigations of attention checks or, more broadly, satisficing behaviors in online environments. We know that attention checks work well for certain task paradigms, but it is not clear *why* they are required in the first place. Perhaps it is simply that without a supervisor looking over their shoulder, workers seek to minimize the effort that they expend. While this is undoubtedly the case for some workers, it might also be that for the majority of workers, variation in working environments and habits is at least partly responsible.

2.2 Multitasking behavior in crowdworkers

One potential cause of inattentiveness in crowdworkers might be multitasking behavior. Multitasking behavior is common in most workplace settings [e.g., Mark et al. 2012; Grundgeiger et al. 2010; Iqbal and Horvitz 2007]. Given that crowdworkers generally do not have bosses or experimenters looking over their shoulders, some workers might be tempted to interleave multiple activities with their work. Currently we know little about the prevalence of such multitasking behavior or the effect it might have on performance. There is some evidence to suggest that having workers take breaks between tasks improves their performance [Rzeszotarski et al. 2013; Dai et al. 2015] and that breaking tasks into smaller chunks improves attentiveness [Cheng et al. 2015]. Overall though, we do not have a good understanding of when and how often people multitask during crowdwork, nor do we know what impact it might have on performance. If multitasking is prevalent and deleterious to performance, can crowdworkers' multitasking behavior be influenced by timely interventions? Developing this understanding is critical, particularly as crowdwork grows to include longer, more cognitively taxing work like graphic design [e.g., Araujo 2013] or software engineering [e.g., K. Mao et al. 2013].

The working environment of crowdworkers differs in important ways from the traditional workplaces that have largely been the focus of multitasking research. At a basic level, the process of working in crowdworking settings is unlike that of traditional working environments [Kittur et al. 2013]. For instance, a task set by one requester is unlikely to have any relationship to the next task, which will often have been set by a different requester. Crowdworkers also spend a lot of time working on repetitive microtasks such as transcription and captioning [e.g., Sampath et al. 2014; Lasecki and Bigham 2012]. These tasks are not always engaging and we know from office-based studies that bored workers tend to be more easily distracted [Mark, Iqbal, et al. 2014]. Physical and social factors also play a significant role in how people multitask in traditional workplace settings [Harr and Kaptelinin 2007; de Vries et al. 2013; Avrahami et al. 2007]. These factors are likely to manifest differently in crowdworking settings where people are distributed and anonymous. This makes an empirical investigation of online multitasking behavior valuable.

We use the 'wildness' afforded by online experimentation to explore the important topic of how multitasking impacts on the quality of work produced through online crowdsourcing platforms. We also investigate whether an intervention that responds

to workers' multitasking behavior is effective in dissuading participants from allowing themselves to be interrupted by other tasks and activities.

2.3 Recording online multitasking behavior

Studying multitasking behavior among crowdworkers is challenging. Leaving the lab means that we sacrifice experimental control. The benefit is that participants in crowds are more likely to be able to engage in the kind of natural multitasking behavior that cannot be easily mimicked in the lab. The drawback is that we have little access to the context of where and how crowdworkers are choosing to complete tasks.

Window-switching trackers have been used to great effect in observational studies to understand the working patterns of traditional office-workers [e.g., Mark et al. 2012; Mark, Wang, et al. 2014; Mark, Iqbal, et al. 2014; Iqbal and Horvitz 2007]. However, these techniques cannot easily be deployed in crowdsourcing platforms: to deeply inspect task-level switching behavior across applications, software has to be installed on participants' machines [although see Rzeszotarski and Kittur 2012 for the possiblities of tracking in-browser activity]. Participants who have to install tracking software may act differently, knowing that they are being monitored. Desktop software also presents logistically insurmountable issues for studying distributed workers using their own devices. The computing environment of crowdworkers is far more heterogeneous than in an office, making developing, installing and supporting desktop software more difficult [Rzeszotarski and Kittur 2012]. One of the advantages of crowd-based studies is that large samples can be obtained with little difficulty. This advantage is compromised if each participant has to go through the process of installing desktop software.

To overcome the challenges of using window logging software, we explore alternative methods for observing multitasking behavior in crowdsourcing settings. We develop a method that tracks participants' multitasking behavior using only the browser window that hosts our task. This is done using a combination of direct window switching measures, indirect activity metrics and pop-up probes that give insight into multitasking behavior. This provides a lightweight and relatively unobtrusive way of inspecting multitasking habits in an online environment.

We investigated the multitasking behavior of crowdworkers. The study asked participants to complete a routine data-entry task that has been widely used to study interruptions and multitasking behaviour [Andrews et al. 2009; Brumby et al. 2013; Gould, Cox, et al. 2013; Li et al. 2006; Ratwani et al. 2008; Trafton et al. 2011]. We monitored our remote crowdworkers to see when they switched away from the task we set them in order to do something else. In this way we were able to record natural multitasking behavior and track its downstream effects on behavior.

3 RESEARCH QUESTIONS

We have three main research questions. Our first research question is: what is the frequency and duration of task switches during online crowdwork, and what activities are people switching to? It is likely that multitasking will be common during online work, just as it is a normal part of many kinds of work [González and Mark 2004; Mark et al. 2005; Chisholm et al. 2000; Westbrook et al. 2010; Loukopoulos et al. 2001]. But understanding the particular characteristics of task-switching during crowdwork is a necessary first step toward developing systems that better support interrupted work. The first condition of our experiment, the control *none* condition, allows us to inspect the baseline frequency and duration of task switches. Our second condition, *solicit*, uses an on-screen probe to gain an understanding of the tasks workers switch to.

Our second research question is: how does multitasking affect performance in crowdworking settings? While it is well known that enforced, experimenter-generated interruptions are deleterious to performance [Monk et al. 2004; Werner et al. 2011], we do not know whether the negative effects of task-switching are quite so clear when

people have discretion over the tasks they complete and the ways in which they manage them. In some circumstances multitasking can have positive effects [Jin and Dabbish 2009; Adler and Benbunan-Fich 2012] by, for example, giving people a chance to rest [Rzeszotarski et al. 2013; Dai et al. 2015]. The experiment presented here provides the first insight into the effects of natural task-switching behavior on task performance.

Our third research question is: if task-switching is frequent and disruptive, is there anything we can do to mitigate its effects? Previous work in crowdsourcing has tried to minimize satisficing behavior in surveys [Kapelner and Chandler 2010], but we do not know whether similar interventions can influence multitasking behavior. The third condition of our experiment, *dissuade*, tests a behavior-sensitive intervention to see if workers' propensity to switch to other tasks and activities can be reduced.

4 METHOD

4.1 Participants

A total of 120 participants (55 female) with a mean age of 33 years (*SD*=10 years; range, 19-72 years) were recruited through the Amazon Mechanical Turk (AMT) crowdsourcing platform. Requirements were set so that workers had to have a task approval rate of 90% or greater and needed to have completed at least 50 tasks before they could participate. The sample was notionally restricted to workers on AMT from the United States.

Conditions were run across sixteen independent HITs ('Human Intelligence Task' – AMT assignments). HITs were posted with alternating conditions until the sampling process was complete. HITs were launched serially; only one HIT was run at a time. The whole study from start to finish was a single HIT assignment from the perspective of the AMT workers. Participants were recruited over the course of three weekdays to ensure as diverse a sample as possible. HITs were completed between 02.00 and 18.00, Eastern Daylight Time.

Participants could not take part in the study more than once. This was enforced by telling workers, in boldface and in the first line of the instructions, that they should not participate more than once. Participants were informed they would not be remunerated for completing the task more than once. WorkerIDs (AMT unique identifiers) were recorded for the purpose of detecting repeat participation. One participant completed the task twice. The data from their second participation were discarded.

Although previous work online [Mason and Watts 2010] and in laboratories [Camerer and Hogarth 1999] suggests that financial incentives have a complex relationship with performance, to eliminate any possibility of remuneration confounding our results, we paid workers USD$6 for completing the 45-minute task.

4.2 Design

The experiment used a between-subjects design with 40 participants in each condition. There was a single independent variable in the experiment, *resumption dialog*. This determined how the task responded to participants switching away from the experiment. It had three levels: *none*, *solicit*, and *dissuade*.

The *none* condition acted as a control; when participants resumed after switching away from the experiment, the switch was recorded by the task without explicitly informing participants. This allowed us to collect baseline data on participants' multitasking behavior.

In the *solicit* condition, participants were asked what activities they had been switching to after they returned to our task. The purpose of this condition was to give insight into the kinds of tasks participants switch to. The prompt that appeared encouraged participants not to alter their multitasking behavior.

In the *dissuade* condition, participants were reminded to focus on the task at hand when they returned to the task. The purpose of this condition was to see if the prevalence of multitasking behavior could be reduced by asking participants not to switch to other tasks and activities.

The primary measures in the *none* and *dissuade* conditions were the number and duration of switches participants made during the experiment. Switching behavior was recorded by binding a function to JavaScript's *blur* and *focus* events in participants' browsers. These were triggered whenever a participant switched away from our task, whether to a different browser tab or to a different application. Timestamps were recorded when participants left and returned so that switch duration could be calculated. In the *solicit* condition, our primary measure of interest was participants' reports of the tasks they were attending to.

As well as recording switching information, we recorded keystroke and trial duration data and noted the number of errors that participants made. These are used to understand general task performance and for a post-hoc analysis of periods of inactivity. This data gives insight into the effectiveness of our window-switching measure.

4.3 Materials

The task used in this experiment was the *Pharmacy Task*, an adaptation of the *Doughnut Machine* [Li et al. 2006]. The *Doughnut Machine* is a routine data-entry task that has been used to investigate the effects of interruptions on performance [Back et al. 2010; Li et al. 2008; Gould, Brumby, et al. 2013].

Participants were given a set of three 'prescriptions' that contained values that they had to copy into one of the five subtasks that make up the task. The subtasks have to be completed in a strict order from left-to-right and top-to-bottom: *Type, Shape, Color, Packaging, Label* (see Figure 1). Participants completed one subtask at a time. After entering the values for a particular subtask, they clicked the 'OK' button that was at the bottom of each subtask. Participants then started working on the next subtask. If participants made any errors while entering the values, or forgot to enter a value, participants were alerted to their mistake. Errors had to be corrected before participants could proceed to the next subtask.

All trials began with an empty *Prescription Sheet* in the middle of the task. To start a trial participants clicked the *New Prescription* button. The button was then replaced with three new orders that had to be entered into the five subtasks. Once all five subtasks had been completed correctly participants clicked the *Process* button. This completed the trial, and the *Prescription Sheet* was again replaced with a *New Prescription* button. To complete a particular subtask, values had to be copied from the *Prescription Sheet* at the center of the screen into the correct fields of each subtask. Subtasks were completed in serial order starting with *Type*.

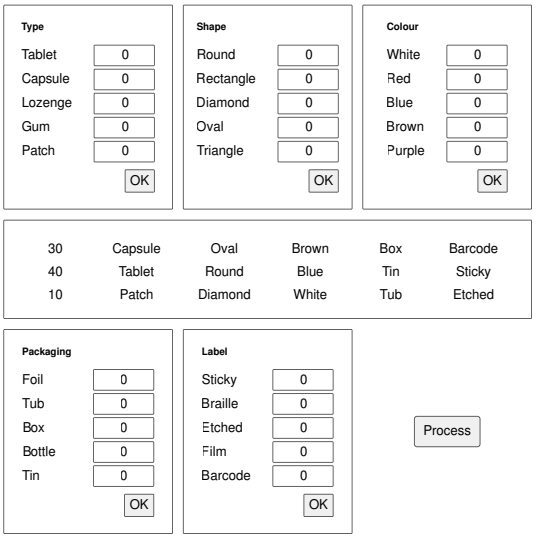

Fig. 1. The Pharmacy Task, as rendered in the browser. Subtasks are completed left-to-right, top-to-bottom: *Type, Shape, Colour Packaging, Label*.

In the example in Figure 1, a participant would have to copy the value 30 from the *Prescription Sheet* into the *Capsule* subtask element of the *Type* subtask. This would be followed by the value 40 into the *Tablet* subtask element. Finally, the value 10 would be copied into the *Patch* subtask element. After clicking 'OK', the participant would then repeat the procedure for the *Shape* subtask. First the participant would copy the value 30 into the *Oval* subtask element of the *Shape* subtask. Then they would enter 40 into the *Round* subtask element. Transcription would continue in this manner for all five subtasks. Although the order of subtasks (*Type, Shape, Colour, Packaging, Label)* had to be completed in serial order, no restriction was put on the order in which subtask elements (e.g., *Tablet*, *Capsule*, *Lozenge*, *Gum*, *Patch* in the *Type* subtask) could be completed. In the example in Figure 1 participants could first enter *40 Tablet,* or *30 Capsule*, or *10 Patch.* The second and third values for each subtask could also be entered in any order.

In addition to the main task, we implemented two kinds of dialog that appeared only when participants switched back to the primary task after an interruption. They only appeared during experimental trials. For the *solicit* condition we developed a response-soliciting dialog (Figure 2). On returning to the task, participants were presented with a grey overlay with three sections. Each section covered a different area of activity that might generate distractions for participants. When the dialog appeared, participants clicked one of six responses in the area that best described their activity. The first section related to activities done on a computer and included web browsing, checking emails and using social networks. The second section related to activities done on phones, including making calls and texting. There was a third section that had non-digital activities listed, such as talking to someone or making a drink. In all three sections there was a 'Doing something else' option, in the event that no option fitted the participant's activity. The purpose of this dialog was to find out what participants were doing while they were gone, so we did not want to put participants off switching. To this end, there was no message reminding participants to complete the experimenting uninterrupted, as there was in the *dissuade* condition. Instead, participants were told that multitasking was 'fine'. We opted against giving participants the option to type a response because it would have made the dialog more disruptive.

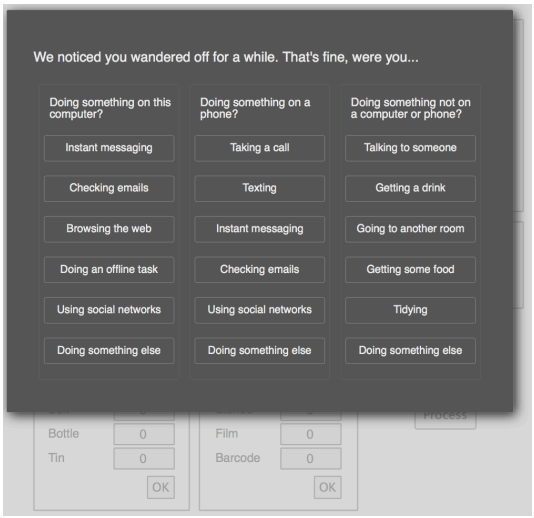

Fig. 2. The dialog in the *solicit* condition asked participants what they had been doing while they were gone

In the *dissuade* condition a dialog (Figure 3) was presented at the very moment a participant returned to the task after an interruption. The intervention consisted of a red overlay that told participants how long they had been gone for and reminded them to complete the experiment uninterrupted. Participants dismissed the dialog in their own time by clicking the 'OK' button. The purpose of this intervention was to remind participants that they should not be switching and to encourage them to stay focused.

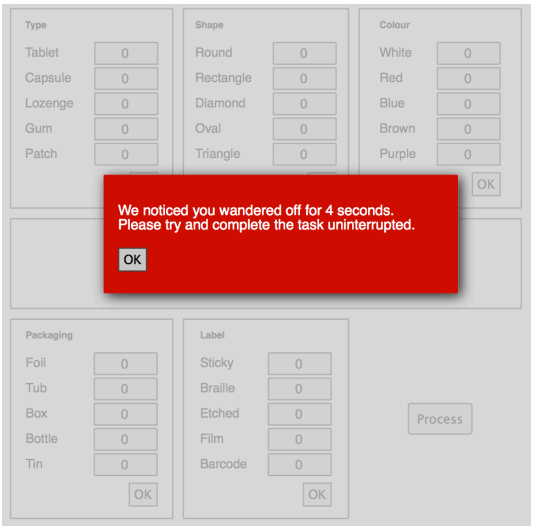

Fig. 3. The timely reminder presented to participants when they returned to the experiment in the *dissuade* condition

4.4 Procedure

The study was listed on Mechanical Turk with the selection criteria described previously. Eligible workers were able to view all introductory materials before they accepted the task. The experiment was opened on a new tab. To aid identification of the tab, the tab under which the experiment was running had a distinctive favicon.

Participants had a page of information about online participation, a page of information about the task, and an introductory video that demonstrated the task. On the page where workers were given guidelines about online working workers were told that "Not following [the instructions] may affect your payment at the end of the study." At the end of the page workers were also told that: *"The most important thing is that*

you are able to set aside around 45 minutes to complete the study uninterrupted." No direct connection was drawn between remuneration and multitasking behavior. Workers were asked to complete the task on a desktop or laptop computer, rather than a phone or tablet. The introductory video had both voice narration and subtitles. Workers who decided to participate completed two training trials to familiarize themselves with the task. The training trials were the same for all conditions: no dialogs appeared after task-switches.

After completing the training trials, a dialog invited participants to proceed to the experimental trials at their own discretion. All participants completed a total of twenty experimental trials. After completing ten trials, participants were given up to three minutes to rest. To reduce the possibility of participants over-exerting their wrists, participants were required to rest for at least sixty seconds before they could move on.

After completing the study, participants completed an eleven-item debriefing survey that asked about their experience of the study. Participants were then given a written debriefing and given a unique code that they entered into Mechanical Turk to confirm that they had completed the task. Participation was approved after participants had finished the experiment. Participants were given an email address to contact if they had questions, but could also contact the experimenters through the AMT interface.

Participants were told that the experiment would last for 45 minutes in the pre-acceptance introductory materials. This period of time included the time taken to work through the introduction, training trials, experiment, questionnaire and debriefing. Participants were given a window of 120 minutes in which to complete the experiment. Before accepting the HIT participants were informed that partial credit was not available. Participants were not informed which condition they had been allocated to before they started. They received no information about the presence of the post-interruption dialogs in the introductory materials: their first encounter with the dialogs was after switching tasks.

5 RESULTS

5.1 Task-switching frequency

We tracked the occasions that people switched to other tasks from the moment they loaded the page until they left the task. Everything from the page loading through to the start of the training trials comprised the *Preamble*. The *Training* trials came after the preamble. Between the training trials and the experimental trials was a *Dialog* screen that informed participants that the *Experiment* was about to start. Half way through the experiment was a *Rest* period. At the end of the experiment was a post-experiment questionnaire and a *Debrief*. A breakdown of the prevalence of switches during each phase is given in Table 1.

Across all participants and all stages of the study there were a total of 933 switches. Of the 120 participants, six (5%) participants went from start to finish with no measurable switches. Most interruptions came during the *Preamble* (454), followed by the *Experiment* (227). These switches were measured directly using our window- and tab-switching technique. As the effects of the *solicit* and *dissuade* conditions were only visible during experimental trials, the *Experiment* phase of the study is of most interest and is therefore the focus of our analysis below.

Task-switching data from the *Experiment* phase of the study showed that participants in the control *none* condition switched most frequently: between 0 and 22 times (M=3.0, SD=5.1). Sixteen participants (40%) in the *none* condition did not switch at all during the experimental trials. The other 60% of participants who did switch averaged five switches during the experimental trials (M=5.0, SD=5.8). Participants in the *dissuade* switched the least frequently: between 0 and 5 times (M=1.0, SD=1.4). Participants in the *dissuade* condition who switched at least once did so less than twice

during the experimental trials, on average (M=1.8, SD=1.3). In the *solicit* condition the number of switches that participants made ranged between 0 and 20 (M=1.7, SD=3.5). For the participants in the *solicit* condition who switched at least once, the average number of switches during the experimental trials was three (M=3.4, SD=4.2). The modal number of switches for all conditions was zero, and the median number of switches was one, indicating large individual and group-level variations in the propensity of participants to switch.

One of the goals of this experiment was to understand the effects of our dialogs on participants' multitasking behavior. For the *dissuade* condition, we hoped that the intervention would lead to a reduction in multitasking behavior. Conversely, we did not intend for the *solicit* condition to put participants off switching; so we also wanted to check if the appearance of the dialog in the *solicit* condition affected participants' multitasking propensity.

	None		**Solicit**		**Dissuade**	
Study phase	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*
Preamble	3.7	(2.8)	3.9	(4.6)	3.8	(3.1)
Training	0.5	(0.9)	0.8	(1.6)	0.3	(0.6)
Dialog	0.1	(0.5)	0.1	(0.2)	0.1	(0.3)
Experiment	3.0	(5.1)	1.7	(3.5)	1.0	(1.4)
Rest	1.3	(1.3)	1.6	(2.3)	1.3	(1.2)
Debrief	0.2	(0.6)	0.2	(0.4)	0.2	(0.5)
Total	8.7	(7.9)	8.1	(8.3)	6.5	(3.8)

Table 1. Mean switch frequency of participants across conditions for each phase of participation. Switches during experimental trials are in bold.

For statistical analysis, a one-way ANOVA was used to determine whether there was a statistically significant effect of dialog type (*none*, *dissuade* or *solicit*) on the frequency of switching. The test revealed a significant main effect $F(2,117)=3.10$, $p<.05$, $\eta_P^2=.05$. Post-hoc tests were used to examine differences between conditions. A t-test comparing the *dissuade* and *none* conditions showed there was a significant difference in switching frequency, $t(79)=2.40$, $p_{adj}=.042$. A comparison of the *solicit* and *none* conditions revealed no significant differences, $t(79)=1.36$, $p_{adj}=.24$. TukeyHSD-adjusted p-values were used to account for multiple comparisons.

The results of the comparison show that the *dissuade* intervention was effective in convincing participants not to interrupt themselves compared to the control *none* condition, where no dialogs appeared. Although participants in the *solicit* condition made fewer switches than participants in the control *none* condition, there was no statistically significant evidence to suggest that participants in the *solicit* condition were discouraged from task-switching.

The study recruited different participants to each condition in a between-subjects design. Might participants in the *dissuade* condition simply have been less likely, as a group, to switch? To be sure that the effects observed were not the product of our sampling, we looked at switch data that were recorded before the experiment started and thus before our manipulations could have had any effect. To do this we looked at switches during the *Preamble* period (see Table 1). This comprises the time between the study loading and participants starting work on the training trials. The frequency of switching during this period gives insight into the 'resting rate' of self-interruption in the sample: the natural frequency of switching. We used a one-way ANOVA to test whether the participants in each condition could be distinguished based on their interruption rates during the *Preamble* (i.e., their 'resting rates' of interruption). The test revealed no significant effect, $F(2,117)=0.05$, $p=.95$. Thus, there was no evidence

to suggest that differences in interruption frequency during the experimental trials were due to differences stemming from the between-subjects design.

5.2 Switch timing

We know that the workers in our study switched often. When did they choose to switch? The timing of switches has been of significant interest to the multitasking and interruptions community [Horvitz et al. 2005; Iqbal and Bailey 2008; Salvucci and Bogunovich 2010]. This interest has been spurred by evidence that suggests that switches are less disruptive at the moment between one subtask finishing and another starting, that is when workload is lower [Bailey and Iqbal 2008; Iqbal and Bailey 2005]. We were interested to know whether participants' switching behavior implied that they were employing the strategy of attending to interruptions at subtask boundaries.

A total of 120 switches were broadly classified as being *between-trials*; *between-subtasks* and *within-subtasks*. Between-trial switches occurred between one trial finishing and the next starting. A total of 45 switches (38%) were of this type. Between-subtask switches occurred between one subtask (e.g., the *Type* subtask) being completed and the next (e.g., the *Shape* subtask) being started. A total of 47 switches (39%) were of this type. Within-subtask switches were deemed to occur anywhere between the start and end of a particular subtask. A total of 28 switches (23%) were of this type. These data are summarized in Table 2.

Switch Timing			
Between-Trial	Between-Task	Within-Task	Total
45	47	28	120

Table 2. Aggregate count of switches at various points in task execution for the 40 participants in the *none* (i.e., no dialog) condition.

We were interested to know what the distribution of these switches was over the period of a trial. Were they evenly distributed or were some moments favored over others? We were looking to see if participants deferred switches until natural breakpoints in the task, as they have been shown to do in previous work [Bailey and Iqbal 2008; Iqbal and Bailey 2005; Salvucci and Bogunovich 2010]. To explore the distribution, we took the 75 switches (47 between-task, 28 within-task) that occurred during trials. We computed a value that represented how far through the trial a switch came. A switch towards the start of a trial had a value closer to zero. A switch toward the end of the trial had a value of one.

We constructed a distribution representing even distances between switches (i.e., what the observed distribution would have looked like if switches were distributed evenly throughout the trials). We took the lowest and highest values from the observed distribution and used them as floors and ceilings in an even-distance distribution that modelled a constant rate. A Kolmogorov-Smirnov test was used to assess the extent to which the observed distribution differed from the even-distance distribution. The test revealed that the observed distribution was significantly different from the even-distance distribution (D=0.23, p<.05).

The distribution (Figure 4) tells us something about the pattern of switches during trials. The positive skew of the distribution clearly demonstrates that participants often switched at the start of trials and towards the end of trials. This suggests that participants were exercising a degree of strategic control over switches, preferring to switch at (or before) the start of trials.

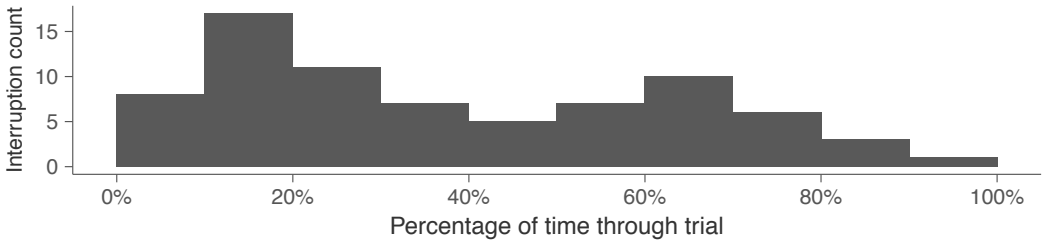

Fig. 4. Distribution of switch timing over proportion of progress through trials.

5.3 Switch duration

As switches during the *Preamble* could not impact task performance, we were most interested in the durations of the switches that occurred during experimental trials. To compute the duration of these switches, we first ignored any participants who made no switches. Including participants with an average switch duration of 'zero' seconds would artificially lower the average duration of switches. This would distort our expectations for the duration of a switch once it had been initiated. For participants who switched a number of times, we used the mean of their switch durations. The resulting means across all participants are shown for each condition in Table 3. There was considerable variation across participants and conditions. The shortest switch that anyone made was 220-ms and the longest was over six minutes. The mean switch duration over the course of the whole study and across all conditions was 39-s (*SD*=60-s). Over the course of the experimental trials, the mean switch duration was 17-s (*SD*=45-s). These data are summarized in Table 3. Participants returned after all switches. No participants got distracted and abandoned work on the task entirely.

Our dialogs in the *solicit* and *dissuade* conditions appeared after every switch. Three short switches brought up three dialogs, but a single, long switch would generate only one dialog. As the appearance of the dialogs was tied to the frequency of switches, we did not expect the dialogs to affect switch duration. Furthermore, the duration of switches is determined to a large extent by the nature of the switched-to task. We had no control over the tasks that participants switched to. Once someone had made the decision to switch to a particular task, their return to our task (and therefore the duration of the switch) would have been dictated by the requirements of the task they switched to, rather than the knowledge they would be faced with a pop-up dialog on their return.

	None		Solicit		Dissuade	
Study phase	M	(SD)	M	(SD)	M	(SD)
Experiment	18-s	(33-s)	10-s	(19-s)	23-s	(69-s)
Overall	39-s	(55-s)	28-s	(30-s)	49-s	(83-s)

Table 3. Mean switch duration for switches during experimental trials in each condition (*Experiment*) and for all switches from *Preamble* to *Debrief* (*Overall*). All values are in seconds.

A one-way ANOVA was performed on the switch duration data from the experimental trials. There was positive skewness to the data because of the prevalence of short switches. A Shapiro-Wilk test suggested that the duration data were not normally distributed, W=.35, p <.001, as a result the data were log-transformed. The ANOVA analysis performed on these log-transformed data found no significant effect of condition on switch durations, $F(2,64)=1.56$, $p=.22$, $\eta_P^2=.05$. We retained outliers in both sets of data because there is a meaningful relationship between the duration of an interruption and its disruptiveness [Monk et al. 2008].

5.4 The effect of switches on task performance

So far, we have demonstrated that the introduction of a behavior-sensitive and timely intervention can be effective in dissuading people from switching to other tasks during

online crowdwork. But is task-switching a problem? Do switches need to be stopped? In this analysis we investigate the impact of switches on performance.

The Pharmacy Task lacks a specific measure of performance like points or number of steps so we used mean trial time as a measure of performance. This was the average time taken by a participant to complete a given trial and excluded the time spent on the introduction, training and debriefing.

Participants who frequently switched would take longer to complete trials than participants who focused on getting through the task undisturbed because of the time spent away from our task attending to other tasks or activities. More interesting is whether switches have a residual impact on performance beyond simply making things take longer – i.e., are they disruptive?

To isolate this information we subtracted the time spent on task-switches during each trial from the total duration of each trial. Each participant completed twenty trials. For each trial we subtracted the time spent on task-switches during that particular trial. If, for example, a trial took 86 seconds, but there was a switch of 18 seconds in the middle of it, the adjusted trial duration would be 68 seconds. Trials that were interrupted several times were adjusted accordingly. After subtracting the time spent on task-switches during each trial from that same trial's total duration, we computed an adjusted average trial time for each participant. Removing the time spent on task-switches eliminated the effect of spending time working on other tasks but, critically, retained any extra time costs associated with recovering after the task was interrupted.

With an adjusted mean trial time for each of the 120 participants, we were able to build a regression model to attempt to explain how much variation in performance stemmed from participants' tendency to engage in task-switching during the study. Average time trial is a useful measure of performance, but is heavily influenced by other factors, such as natural differences in participants' working speeds. As interaction with the task involved a significant volume of typing, we believed that typing speed would explain a significant portion of individuals' working speeds and provide a point of comparison for the influence of task-switches on performance. Inter-keystroke interval (IKI), the period between two sequential keystrokes, was included in the model.

Another factor that may have influenced mean trial time was errors. The Pharmacy Task required subtasks to be completed in a specific order, and for each subtask to be completed correctly. Any variations to the correct order were logged as an error and participants had to rectify them before continuing. Fixing errors adds to the average trial time. Participants made a mean of four errors over the course of the experiment (SD=3) with a mode of two. The total number of errors made during the experiment varied between zero and twenty-five. The distribution of errors is illustrated in Figure 5. A one-way ANOVA showed there was no significant effect of dialog condition on error rate ($F(2,117)=0.20$, $p=.81$), suggesting that the dialogs did not interfere with participants' task completion accuracy. We also considered error rate because it is likely to be indicative of both participants' understanding of the task as well as their fastidiousness in completing it.

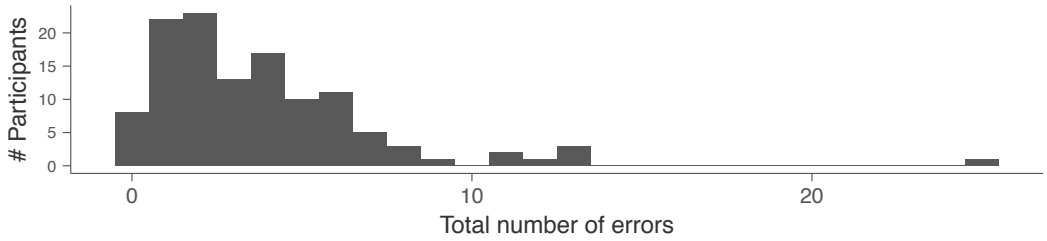

Fig. 5. Histogram of total errors made by participants during experimental trials. The data are positively skewed, with a mean of four errors and a mode of two errors.

We built a one-step linear regression model to apportion the variance in participants' trial times. The first predictor in the model was the total number of switches that a participant made during the experimental trials. Each task-switch has costs associated with it. These costs accumulate across the experiment with each switch. We wanted to know whether they had a material effect on performance. The second predictor added to the model was switch duration. Longer interruptions are known to be more disruptive than shorter ones [Monk et al. 2008]. To maintain commensurability with other time-based measures in the model we used raw average task-switch durations without transformation. The third predictor was inter-keystroke interval (IKI). IKI was computed using the time between consecutive strokes within a single string. Gaps between entering the last character of one string and the first of the next were ignored. Participants' mean IKIs were calculated by averaging all of their IKIs over the experiment. This measure gives an indication of typing speed. The fourth predictor added to the model was total number of errors that a participant made. A summary of the predictors is given in Table 4.

Predictor	Details
Total switches	The total number of task switches for each participant during trials. (Individual frequency.)
Mean switch duration	The average duration of switches made by a participant. (Individual mean.)
Inter-keystroke interval	The average time between two consecutive keystrokes computed for each participant. (Individual mean.)
Total errors	The total number of errors made by each participant during task execution. (Individual frequency.)

Table 4. Regression predictors as entered in the one-step model.

The linear regression model explained a significant amount of variation in participants' mean trial time (R^2_{adj}=.37, $F(4,115)=18.76$, $p<.001$). The assumption of independent errors was met (Durbin-Watson = 2.01). The minimum standardized residual was -1.8. The maximum was 5.0 (SD=0.98). Three cases (2.5%) were ±2 standard deviations from the mean, indicating the model was not overly biased (see Figure 6).

Predictor	*B*	SE *B*	β	*t*	*p*
Total switches	1150	533	.16	2.16	.03*
Mean switch duration	0.09	0.06	.12	1.58	.12
Inter-keystroke interval	79	12	.49	6.71	<.001*
Total errors	2239	560	.29	4.00	<.001*

Table 5. Breakdown of predictors for regression model that predicts total trial time. Asterisk denotes predictors significant at the $p<0.05$ level. ***B***s are unstandardized regression coefficients in milliseconds and **SE *B***s are standard errors of those coefficients. **β**s are standardize coefficients. **t**s are t-statistics and ***p***s are p-values for those statistics.

As can be seen in Table 5, three predictors explained significant variation in the model. The largest share of the variation was explained by participants' typing speed. The faster they typed, the quicker they finished each trial. Error rate explained the second largest portion of the variance. The more errors participants made, the longer it took them to complete trials. For the switching-related predictors the effects were smaller. The frequency at which participants switched to other tasks explained a smaller portion of variation, but still had a significant effect on mean trial time. The more participants switched, the longer they took on average to complete a trial – even when the time spent on task-switches was accounted for. The model estimates that, for this task, each switch adds 1150-ms to the average trial time. Given that an average trial takes approximately 70 seconds, the model suggests that each task-switch

increases the duration of an average trial by around 1.6%. This excludes any time actually spent on other tasks; this is the isolated effect of switch costs. There was a trend for longer switches to increase average trial time (after accounting for the time spent on switches), but this was not significant.

This model provides insight into the likely effect of task-switches on average performance. That some trials are interrupted is accounted for by the model. This means the model is useful for understanding the likely effects of task-switches on a task completed by randomly sampled AMT workers; some workers will switch, some workers will not. Workers will switch during some trials, but not during others.

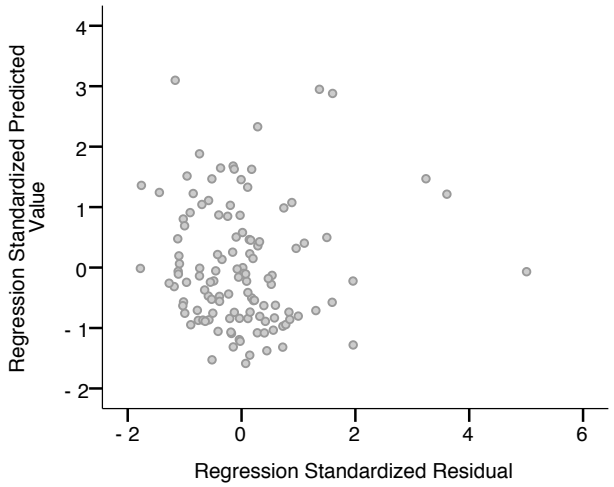

Fig. 6. Standardized residuals from regression model. Note three outliers ±2 standard deviations from the mean standardized residual.

5.5 Pauses in activity

The previous analyses were based on data obtained by monitoring browser *focus* and *blur* events. This meant that task switches outside the browser could not be captured (i.e., if participants switched to an activity away from their computer). To gain some insight into these extra task switches we examined keystroke data, looking for long pauses in task activity.

Inter-keystroke interval (IKI) data provides a measure of typing speed. As typing speed is relatively consistent, unusually large intervals between one keystroke and the next when typing a string might suggest that participants had switched to another activity. The mean IKI was 322-ms (*SD*=429-ms). On seven occasions, an IKI was filled by a task-switch that had already been recorded with the window-switching code. These IKIs were discarded from the rest of the IKI data analysis.

The distribution of IKIs is illustrated in Figure 7. The longest interval was greater than 40-s. In Table 6 we break down the IKIs starting with those that are greater than two or more standard deviations from the mean. To give a closer view of the longer IKIs, Figure 8 provides a crop of the right-hand tail of Figure 7. These longer IKIs make up a relatively small proportion of all intervals. In absolute terms, however, there are many occasions when there are long delays between characters. There were 80 occasions when the delay exceeded six standard deviations of the mean (i.e., more than 2902-ms).

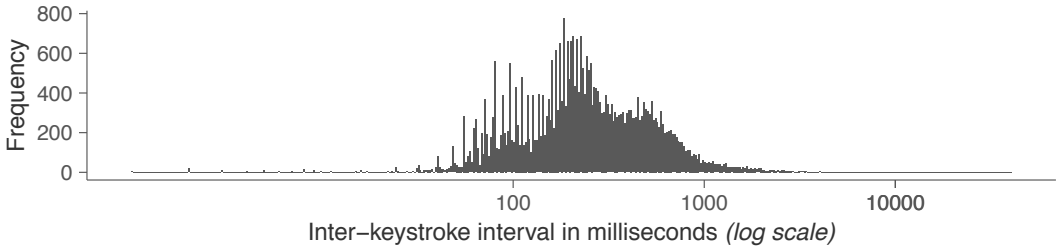

Fig. 7. Distribution of inter-keystroke intervals.

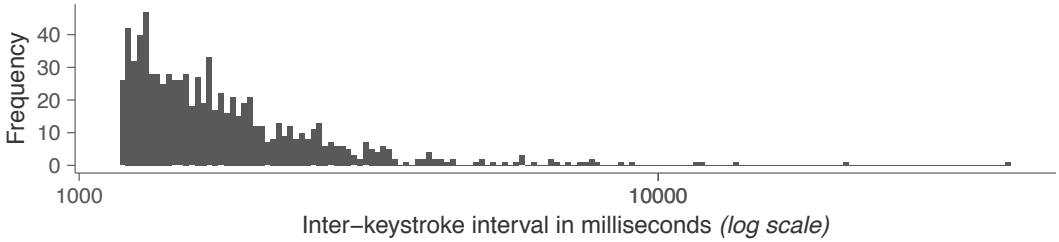

Fig. 8. Distribution of inter-keystroke intervals, with focus on intervals greater than two standard deviations more than the mean (i.e., (2 × 430-ms + 322-ms, 1182-ms).

From mean	Count	Proportion
> +2 SDs, 1182-ms	832	2.2%
> +3 SDs, 1612-ms	415	1.1%
> +4 SDs, 2042-ms	216	0.6%
> +5 SDs, 2472-ms	136	0.4%
> +6 SDs, 2902-ms	80	0.2%

Table 6. Total number and proportion of IKIs at different degrees from the standard deviation.

We next use the IKI data to examine whether participants were more likely to switch activities after completing a subtask (i.e., we look for periods of inactivity at subtask boundaries). To do this we computed the elapsed time between the final action in a subtask (i.e., clicking the 'OK' button) and the first keystroke on the next subtask. We considered only transitions from one of the five subtasks to another of the five subtasks. Thus, there were four between-subtask boundaries in each trial. This yielded a total of 9,600 transitions of interest (four transitions per trial, 20 trials per participant, 120 participants). There were 38 transitions where a task-switch had already been detected by our window-tracking approach. These transitions were discarded from the following analysis.

Across all participants the transition from completing one subtask to starting work on the next was a mean of 4.3-s (*SD*=5.8-s). The distribution of inter-subtask intervals is illustrated in Figure 9. Again, we produced a breakdown of these transitions by how far they deviated from the standard deviation. Figure 10 and Table 7 show the frequency of inter-subtask transitions that were in excess of two standard deviations of the mean. The data revealed 101 inter-subtask intervals of over 15-s in duration (~+2 SDs). Nearly 50 of these were over 21-s in duration (~+3 SDs).

Our primary measure of task-switching – blur and focus events in the browser – revealed 227 switches during the experimental trials. Using the additional measures reported in this section, we found additional moments where it seems likely that participants had switched away from the primary task. For instance, the inter-subtask interval analysis revealed 48 intervals in excess of the mean duration of directly measured switches. If all 48 were task-switches, our total count of switches from the direct measure would be a 21% underestimate. Variations in the inter-keystroke intervals are smaller in absolute terms, but in relative terms, 80 intervals were more than six standard deviations from the mean, suggesting at least a degree of inattention. If all of these longer intervals were indeed switches, it would suggest that our browser-based measure captures around 64% of task-switches. However, as these

additional switches can only be inferred from these inferred measures, and were not captured by our primary measure, we exclude them from any conclusions we draw, except to say that our primary measure is most likely an underestimate of the prevalence of task-switching behavior.

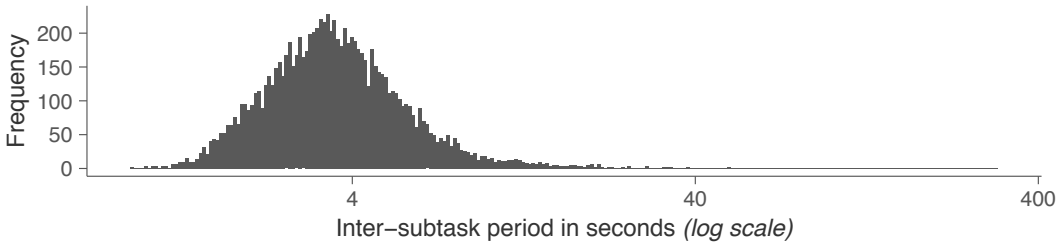

Fig. 9. Distribution of inter-subtask intervals.

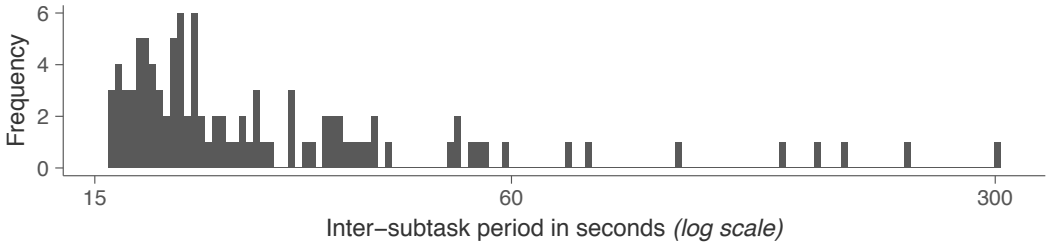

Fig. 10. Distribution of inter-subtask intervals, with focus on intervals greater than two standard deviations (2 × 5826-ms) more than the mean (4258-ms). (i.e., greater than 15,910-ms)

From mean	Count	Proportion
> +2 SDs, 15.9-s	101	1.1%
> +3 SDs, 21.7-s	48	0.5%
> +4 SDs, 27.6-s	33	0.3%
> +5 SDs, 33.4-s	25	0.3%
> +6 SDs, 39.2-s	16	0.2%

Table 7. Total number and proportion of inter-subtask intervals at several degrees of standard deviations from the mean.

5.6 Responses to 'solicit' condition

We wanted to know more about what participants were actually doing when they switched to other tasks. To do this, we asked participants in the *solicit* condition to tell us what they had been doing when they came back to the experiment after switching to other tasks.

Half of the participants in the *solicit* condition did not switch to other tasks during the experimental trials, but we still had data from the other twenty participants who did switch and had to respond to the dialog that appeared when they returned. Between them, these participants switched (and so completed the dialog) a total of 85 times. On some of these occasions, participants seemed to be continuously engaged in another activity and reported doing the same activity on a number of occasions in quick succession (e.g., instant messaging). Where these 'runs' of activity occurred, we collapsed the reports into a single report to prevent a small number of participants from dominating the sample. Removing these runs left thirty 'unique' switches. The breakdown of these switches is given in Table 8.

Of the thirty cases, 10 were computer-oriented, 13 were mobile phone-oriented and the remaining seven were precipitated by something other than a phone or computer. Seven of the 10 computer-oriented switches were to something that was not on our list. Of the switches where a participant did pick an activity from our lists, one person went to check their emails, one person reported instant messaging (evidenced by six further switches during the experiment) and one person reported using a social network.

Phone-based activity was dominated by call-taking; seven of thirteen phone-related switches were for this purpose. One participant reported checking emails on their phone. The other five switches to phone-based tasks did not fit into any of the categories. If the switch was not computer or mobile related, participants were most likely to leave to get a drink (3 of 7). Only one switch was reported to have been the result of someone talking to a participant. In total, 13 of the 30 non-computer, non-phone switches recorded were precipitated by something that we did not have on our lists.

Type	Activity	Frequency
Computer-oriented	Doing something else	7
	Instant messaging	1
	Using social networks	1
	Checking email	1
Mobile-oriented	Taking a call	7
	Doing something else	3
	Checking email	2
Other	Getting a drink	3
	Doing something else	3
	Going to another room	1
	Talking to someone	1

Table 8. Response frequencies from the *solicit* condition

6 DISCUSSION

6.1 Summary of results

After switching, participants in the *none* condition saw no dialog. Participants in the *dissuade* condition saw a dialog that asked them to stop switching. Participants in the *solicit* condition saw a dialog that asked them to indicate the activity they had switched to. The results of the experiment showed that:

- There was a main effect of dialog type on task-switching frequency. Participants switched most often in the *none* condition and least often in the *dissuade* condition. Post-hoc tests showed that the *dissuade* dialog reduced switch frequency compared to the baseline *none* condition. The *solicit* condition did not reduce switch frequency compared to the baseline.
- There was no evidence that the groups differed in their propensity to switch before the intervention dialogs appeared.
- When participants did switch, they tended to switch towards the beginning of the task.
- Dialog type had no effect on how long switches lasted, only on the frequency of their occurrence.
- Frequent switching yielded slower performance, even after accounting for the time spent on task-switches and allowing for individual differences in typing speed and error rate.
- Browser-based monitoring of task-switching captures many instances of multitasking behavior, but activity-logging points to additional out-of-browser task switching.
- Participants switched to a variety of different activities.

6.2 Characterizing multitasking in crowdworking

The results of the experiment revealed that task-switching is prevalent during online experiments, just as it is in traditional workplace settings [e.g., Dabbish et al. 2011; González and Mark 2004]. We found that the majority (60%) of participants switched away from the experimental trials. Ninety-five percent of participants switched to

other activities at some point between loading the study and completing the debriefing. Based on the switches we were able to observe using our direct browser-based measures, participants switched every five minutes on average during the experimental trials. This is comparable to previous work that has shown interruptions to workplace activity occur every three to seven minutes [see González and Mark 2004; Chisholm et al. 2000; Westbrook et al. 2010]. Responses solicited from participants suggest that when they did switch, they switched to a wide variety of tasks both at and away from their computers. Our data suggests that working remotely does not free people from interruptions; the frequency of interruptions is similar, they just come from different sources.

6.3 The effect of task-switching on performance

Should task-switching in crowdworking settings be a cause for concern? After all, we know that breaks can be beneficial to crowdworkers' performance [Rzeszotarski et al. 2013; Dai et al. 2015] as long as they are scheduled in a way that does not interfere with tasks.

We looked at the effect of switches on performance during task execution. Excluding the time participants spent away from the experimental task attending to other tasks, each switch added 1.6% to the average trial time. This supports the idea that there are planning and re-encoding costs associated with interruptions to activity [see Trafton et al. 2011; Brumby et al. 2013; Salvucci et al. 2009]. These costs negatively impact task performance over and above the time spent away from the primary task [Salvucci 2010].

There was no significant relationship between switch frequency and the frequency of errors made during execution of our experimental task. Previous work has often demonstrated that interruptions lead to downstream errors [e.g., Brumby et al. 2013; Monk 2004; Li et al. 2006]. It is likely that the low error rates that we observed were a feature of the task that we used. The cognitive load of the main task was low: participants simply transcribed numbers from one part of the screen to another – no operations had to be performed and no problem state [Borst et al. 2015; Salvucci et al. 2009] had to be maintained. Interruptions to work can induce higher error rates when primary tasks are cognitively burdensome [e.g., Hodgetts and Jones 2006; Gould, Brumby, et al. 2013]. The other factor contributing to the low error rates is likely to have been the confirmation stage of each subtask. If participants did make errors, they could easily be caught and corrected before entries were confirmed. In tasks where errors can be corrected before a confirmation step, we would expect recorded (or 'confirmed') error rates to be relatively lower compared to tasks without confirmation steps.

6.4 Discouraging multitasking behavior

In our experiment we found that switches occurred every five minutes, on average. There were also some indications that the more frequently workers switched, the greater the negative effect on their performance. We discovered that by deploying a timely intervention, workers could be persuaded to switch tasks less frequently. Switch frequency during the experimental trials decreased from an average of three switches in the control *none* condition to one in the *dissuade* condition where the timely intervention appeared.

Is task-switching sufficiently prevalent and disruptive to be worth doing something about? The answer to this question depends, we argue, on setting the costs of implementing an intervention against its potential benefits. Our results show that other factors, such as typing speed, are likely to play a larger role in overall task performance (49% in our model) than task-switches (16% in our model). But making people better typists requires significant training [Keith and Anders 2007], which may have minimal returns anyway, given that AMT workers are already very competent

typists [Crump et al. 2013, p.3]. The intervention we developed in the *dissuade* condition can easily be implemented in any AMT task by tracking standard window events. It is also a targeted intervention that responds to participants' behavior: by only targeting participants who engage in task switching, focused participants do not have their time wasted on completing inappropriate attention-capturing tasks. Incremental improvements in performance obtained at almost zero cost and when aggregated over a market as large as AMT, have the potential to lead to significant absolute productivity gains.

Our task was straightforward to complete, and as such may not have been susceptible to some of the negative effects of multitasking. In more complex tasks that make higher cognitive demands on workers, like graphic design [Araujo 2013] or software development [K. Mao et al. 2013], the negative effects of switching are likely to be exacerbated. In such tasks, the effects of task-switching on performance could be particularly deleterious.

6.5 Broader implications for crowd research

For the broader crowdsourcing community, our findings offer requesters an indication of how often workers switch to other tasks and the duration of their absences. Our results show that crowdworkers, like office workers, cannot be relied upon to work undisturbed, and this should be taken into consideration when evaluating worker performance.

One of the findings from this study that is relevant to all requesters of crowdwork relates to instruction adherence. In the introduction, participants were asked to "set aside approximately 45 minutes to complete the task uninterrupted". If they were interrupted, participants were told, it would "cause problems" for our experiment and that they would be paid less. Yet participants in the study still engaged in task switching behavior. This replicates the findings of previous work that has shown that online participants do not always follow experimenter instructions [Kapelner and Chandler 2010]. As our results show, a timely intervention can be used to remind participants of important instructions at the moment that non-compliant behavior occurs.

For those using the crowdsourcing platforms as a tool for conducting research, our results suggest that caution is required, particularly for studies where time-sensitive measures like reaction times or decision times are dependent variables [Komarov et al. 2013; Gould, Cox, Brumby, et al. 2015]. The occurrence of switches during sensitive periods could skew results. We recommend that researchers monitor and record task-switching behavior as participants work through online studies. Using task switching metrics like those outlined in this paper to augment other performance measures might help to better explain both anomalous and normal performance.

Our results also suggest that requesters posting time-sensitive tasks should be mindful of interruptions, and try to minimize them where possible. This is because our results indicate that task-switches have a broadly negative effect on worker performance. Our regression model predicts a 1.6% increase in average trial duration for each switch a worker makes. A few task-switches can have a large cumulative effect across a whole task.

Our statistical model of task-switching frequency accounts for the fact that some workers switch frequently while others do not switch at all. This means that for workers who are prone to frequent task switching, the model underestimates the likely frequency of switches. Detecting workers who are 'switchers' and trying to nudge them toward staying focused might result in improved performance across tasks. Marginal performance improvements should be of particular interest to workers who are frequent switchers: most tasks on AMT are 'piecework' so workers are paid for what they produce and not the time they spend producing it. Frequent switching increases the amount of potentially productive time lost to switching costs. Helping workers to

minimize these relatively small switching costs could add up to substantial productivity increases over the thousands of tasks that full-time AMT workers complete.

6.6 Implications for our understanding of multitasking

Laboratory-based multitasking research has frequently made claims about the negative effects of multitasking behavior on performance [e.g., Brumby et al. 2013 ; Dabbish and Kraut 2004; Katidioti and Taatgen 2014; Liu et al. 2014; Monk et al. 2004; Salvucci and Bogunovich 2010]. In a typical multitasking experiment, participants are given a task to perform. This task is then occasionally interrupted by a secondary task designed by the experimenter. Enforced interruptions like these disrupt activity, incurring time costs and increasing the likelihood of errors [Bailey and Konstan 2006; Trafton et al. 2011]. Even very short interruptions of a couple of seconds have significant negative effects on performance [Altmann et al. 2013]. Validating the claims that are based on evidence from lab studies has been difficult in practice because measuring 'performance' in most workplaces is difficult [Grundgeiger et al. 2010; Mark et al. 2012].

There is some evidence in our results that performance suffered when participants allowed themselves to be interrupted. This is significant because it suggests that discretionary switching to secondary tasks has negative effects on performance, even when people have free rein over the switches they choose to make. This gives support to the idea that processing interruptions has unavoidable cognitive costs.

Our results also inform our understanding of the moments at which people switch to other tasks. Interruptions that require immediate and non-negotiable attention are atypical of the interruptions we are likely to encounter in our day-to-day lives. For example, many people receive email notifications while they are busy working on other, more important, computer-based activities. Research using scenarios where people have discretion over when to switch tasks has demonstrated that people tend to postpone switching to other tasks until moments of lower workload in their primary task [Bogunovich and Salvucci 2011; Salvucci and Bogunovich 2010; Iqbal and Bailey 2007; Janssen et al. 2012].

Our data suggest that participants made strategic decisions about when to switch to other tasks. Switches were more likely before the start (or after the end) of trials, suggesting a preference for switching at natural breakpoints between trials. Although switches tended to occur at the start and end of trials, our results also show that even with complete discretion over when switches occurred, people were still willing to switch in the middle of a task. This behavior might have occurred because participants' strategies for task-switching are not well developed or because they lack motivation. The most likely explanation is that participants saw other time-critical tasks as being more important and that there was more utility to be had from switching immediately than there was from waiting until a 'convenient' breakpoint at the end of a trial.

The results of our study suggest that this seemingly suboptimal behavior occurs with switches that have real utility to participants. In the lab it is difficult to generate tasks that have sufficient value to participants to induce switching behavior. Working environments, conversely, are awash with distractions that we perceive to be important and that immediately require our attention. With this study we have made progress toward understanding people's multitasking habits in more natural contexts.

6.7 Implications for investigations of multitasking

In this study we embraced the loss of control associated with Internet-based experiments, appropriating it to investigate participants' multitasking habits. Unlike typical multitasking studies that give participants a primary and secondary task to work on, we instead gave participants a single task and observed how they interleaved its execution with other tasks that they themselves perceived to be worth switching to.

For researchers investigating multitasking behavior, we have documented the exploration of multitasking behavior using a novel method and in a new domain. In the past there has been a disconnect in the multitasking literature between laboratory-based research and situated work [Gould et al. 2012; Janssen et al. 2015]. Our work makes an initial contribution to the effort to move lab-like studies into more naturalistic settings.

New ways of working, such as crowdsourcing platforms, should be of interest to interruptions and multitasking researchers because they potentially offer both a research tool and a new working domain to be explored. Crowdsourcing platforms afford intriguing prospects for multitasking research for two reasons. First, they offer the possibility of assigning specific tasks to workers. Going into workplaces and setting people particular tasks to perform is problematic [see Mark et al. 2012]. As a consequence, situated studies end up recording activities for which there may be no easily operationalized definition of 'good performance'. Crowdsourcing platforms obviate this difficulty by giving researchers control of the tasks that workers perform. The second promising feature of crowdsourcing platforms is that crowdwork takes place in environments outside laboratories. This means that as crowdworkers go about their work, they are likely to have to deal with distractions which, unlike the distractions in laboratory studies, are likely to have true utility. Given these properties, crowdsourcing offers an opportunity for naturalistic experimentation that combines the advantages of both situated and lab work. Further exploration of crowdsourcing platforms' potential for hosting naturalistic experiments should be undertaken.

6.8 Limitations

Our results showed that the intervention in the *dissuade* condition reduced participants' propensity to switch to other tasks. But it is not clear from our results *why* the intervention was effective. It is possible that participants were concerned that multitasking might affect their payment at the end of the study; when it became clear that this behavior was being monitored, they adjusted. It is also possible that participants simply found the dialog to be irritating and adjusted their behavior to avoid seeing it. Other design interventions may also have the intended effect of increasing attentiveness: offering the chance to break regularly and at less disruptive moments can also improve performance [Rzeszotarski et al. 2013].

In addition to the *dissuade* condition, which asked participants not to switch to other tasks, another type of post-switch dialog, *solicit*, asked participants which tasks they were switching to. We expected responses to be dominated by computer-based activities because in order for the loss of task focus to be detected, participants had to switch from the experiment to something else on their computer. This limited our ability to detect lapses in attention (although automatic tasks like screensavers or virus scanners would also have triggered a switch event). While switches could be precipitated by non-computer activities, these activities had to result in participants performing some computer-based activity in order for them to be detected.

Another limitation of the method we used is that it does not explicitly account for tab switching activities related to task management in Amazon Mechanical Turk [see, e.g., Lasecki et al. 2015]. Such behavior may have been the cause of some of the switches we measured. Collecting data on these behaviors has provided insight into the working patterns of Turkers. Even very brief switches to check on the status of other assignments might have a detrimental effect on performance. Previous work makes clear that even very short interruptions of less than a second have a deleterious effect on performance [Altmann et al. 2013].

Detecting periods during which participants were away from their computers but with the experiment still in focus on the screen is difficult, because it requires a reliable estimate of participants' average speed and thinking time while working on the task

[Rzeszotarski and Kittur 2011]. Although we were able to infer possible switching behavior from a couple of inactivity metrics, other methods could be employed to give a richer understanding of task switching behavior away from computers. Mouse movements and activity timers might allow researchers to make reasonable assumptions about when participants are no longer focusing on an experiment without interrupting participants who have paused for thought [A. Mao et al. 2013]. A 'dead man's handle' could also be helpful in this regard – users could be prompted to confirm they were working on the task at random intervals, with any responses over a threshold duration counted as inattentiveness. All of these techniques would need to be implemented in a way that does not disturb participants who are actually working. Using these methods might give some insight into how the dialogs in our study affected participants' multitasking behavior. It might be that the dialogs encouraged participants to modify their switching behavior so that they did not move out of the browser window. Using a variety of other attention tracking techniques would give us a better understanding of how and why our *dissuade* intervention actually influenced behavior: it might be that the intervention simply changes the expression of multitasking behavior rather than its prevalence.

7 CONCLUSION

Researchers who have viewed crowdsourcing platforms as a new kind of workplace have discussed the need for these platforms to develop comfortable, meaningful working environments for the ever-increasing number of people earning at least part of their living through crowdwork and other forms of distributed online working [Kittur et al. 2013]. There is a long research tradition that seeks to understand the nature of work and workplaces, whether technically [Shackel 1962] or narratively [Terkel 1974]. While this understanding is well developed in a number of environments, we are only just beginning to understand what work and workplaces look like, and could look like, for crowdworkers. The work we present here makes a contribution to our understanding of multitasking behavior in this nascent workplace and advances a new element for inclusion in the design of this working environment.

8 ACKNOWLEDGEMENTS

This work has benefitted greatly from the editorial direction of Ed Chi and from the input of four anonymous ToCHI reviewers. We are grateful for their professionalism. Jake Rigby and Judith Borghouts helped with the preparation of this manuscript. This work was supported by the UK Engineering and Physical Sciences Research Council grants EP/G059063/1 and EP/L504889/1.

9 REFERENCES

Rachel F. Adler and Raquel Benbunan-Fich. 2012. The effects of positive and negative self-interruptions in discretionary multitasking. In *Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts*. CHI EA '12. New York, NY, USA: ACM, 1763–1768. DOI: http://dx.doi.org/10.1145/2223656.2223706

Erik M. Altmann, Gregory J. Trafton, and David Z. Hambrick. 2013. Momentary Interruptions Can Derail the Train of Thought. *J Exp Psychol Gen* 143, 1 (2013), 215–226. DOI: http://dx.doi.org/10.1037/a0030986

Alyssa Andrews, Raj M. Ratwani, and J. Gregory Trafton. 2009. Recovering From Interruptions: Does Alert Type Matter? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 53, 4 (2009), 409. DOI: http://dx.doi.org/10.1177/154193120905300451

Ricardo Matsumura Araujo. 2013. 99designs: An Analysis of Creative Competition in Crowdsourced Design. In *First AAAI Conference on Human Computation and Crowdsourcing*. AAAI, 17–24.

Daniel Avrahami, James Fogarty, and Scott E. Hudson. 2007. Biases in human estimation of interruptibility: effects and implications for practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07. New York, NY, USA: ACM, 50–60. DOI: http://dx.doi.org/10.1145/1240624.1240632

Jonathan Back, Duncan P. Brumby, and Anna L. Cox. 2010. Locked-out: investigating the effectiveness of system lockouts to reduce errors in routine tasks. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '10. New York, NY, USA: ACM, 3775–3780. DOI: http://dx.doi.org/10.1145/1753846.1754054

Brian P. Bailey and Shamsi T. Iqbal. 2008. Understanding changes in mental workload during execution of goal-directed

tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction* 14 (January 2008), 1–28. DOI: http://dx.doi.org/10.1145/1314683.1314689

Brian P. Bailey and Joseph A. Konstan. 2006. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior* 22, 4 (July 2006), 685–708. DOI: http://dx.doi.org/16/j.chb.2005.12.009

Tara S. Behrend, David J. Sharek, Adam W. Meade, and Eric N. Wiebe. 2011. The viability of crowdsourcing for survey research. *Behav Res Methods* 43, 3 (September 2011), 800–813. DOI: http://dx.doi.org/10.3758/s13428-011-0081-0

Michael S. Bernstein, Joel Brandt, Robert C. Miller, and David R. Karger. 2011. Crowds in Two Seconds: Enabling Realtime Crowd-powered Interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. UIST '11. New York, NY, USA: ACM, 33–42. DOI: http://dx.doi.org/10.1145/2047196.2047201

Peter Bogunovich and Dario D. Salvucci. 2011. The effects of time constraints on user behavior for deferrable interruptions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11. New York, NY, USA: ACM, 3123–3126. DOI: http://dx.doi.org/10.1145/1978942.1979404

Jelmer P. Borst, Niels A. Taatgen, and Hedderik van Rijn. 2015. What Makes Interruptions Disruptive?: A Process-Model Account of the Effects of the Problem State Bottleneck on Task Interruption and Resumption. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. New York, NY, USA: ACM, 2971–2980. DOI: http://dx.doi.org/10.1145/2702123.2702156

Duncan P. Brumby, Anna L. Cox, Jonathan Back, and Sandy J.J. Gould. 2013. Recovering from an interruption: Investigating speed-accuracy trade-offs in task resumption behavior. *J Exp Psychol-Appl* 19, 2 (2013), 95–107. DOI: http://dx.doi.org/10.1037/a0032696

Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's Mechanical Turk A New Source of Inexpensive, Yet High-Quality, Data? *Perspect Psychol Sci* 6, 1 (January 2011), 3–5. DOI: http://dx.doi.org/10.1177/1745691610393980

Colin F. Camerer and Robin M. Hogarth. 1999. The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. *J Risk Uncertainty* 19, 1–3 (December 1999), 7–42. DOI: http://dx.doi.org/10.1023/A:1007850605129

Jesse Chandler, Pam Mueller, and Gabriele Paolacci. 2014. Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behav Res* 46, 1 (March 2014), 112–130. DOI: http://dx.doi.org/10.3758/s13428-013-0365-7

Justin Cheng, Jaime Teevan, Shamsi T. Iqbal, and Michael S. Bernstein. 2015. Break It Down: A Comparison of Macro- and Microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. New York, NY, USA: ACM, 4061–4064. DOI: http://dx.doi.org/10.1145/2702123.2702146

Carey D. Chisholm, Edgar K. Collison, David R. Nelson, and William H. Cordell. 2000. Emergency Department Workplace Interruptions Are Emergency Physicians "Interrupt-driven" and "Multitasking"? *Academic Emergency Medicine* 7, 11 (2000), 1239–1243. DOI: http://dx.doi.org/10.1111/j.1553-2712.2000.tb00469.x

Matthew J.C. Crump, John V. McDonnell, and Todd M. Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE* 8, 3 (March 2013), e57410. DOI: http://dx.doi.org/10.1371/journal.pone.0057410

Laura Dabbish and Robert E. Kraut. 2004. Controlling interruptions: awareness displays and social motivation for coordination. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*. CSCW '04. New York, NY, USA: ACM, 182–191. DOI: http://dx.doi.org/10.1145/1031607.1031638

Laura Dabbish, Gloria Mark, and Víctor M. González. 2011. Why do I keep interrupting myself?: environment, habit and self-interruption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11. New York, NY, USA: ACM, 3127–3130. DOI: http://dx.doi.org/10.1145/1978942.1979405

Peng Dai, Jeffrey M. Rzeszotarski, Praveen Paritosh, and Ed H. Chi. 2015. And Now for Something Completely Different: Improving Crowdsourcing Workflows with Micro-Diversions. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '15. New York, NY, USA: ACM, 628–638. DOI: http://dx.doi.org/10.1145/2675133.2675260

Frédéric Dandurand, Thomas R. Shultz, and Kristine H. Onishi. 2008. Comparing online and lab methods in a problem-solving experiment. *Behav Res* 40, 2 (May 2008), 428–434. DOI: http://dx.doi.org/10.3758/BRM.40.2.428

Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are your participants gaming the system?: screening mechanical turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '10. New York, NY, USA: ACM, 2399–2402. DOI: http://dx.doi.org/10.1145/1753326.1753688

Laura Germine, Ken Nakayama, Bradley C. Duchaine, Christopher F. Chabris, Garga Chatterjee, and Jeremy B. Wilmer. 2012. Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychon Bull Rev* 19, 5 (October 2012), 847–857. DOI: http://dx.doi.org/10.3758/s13423-012-0296-9

Victor M. González and Gloria Mark. 2004. "Constant, constant, multi-tasking craziness": managing multiple working spheres. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '04. New York, NY, USA: ACM, 113–120. DOI: http://dx.doi.org/10.1145/985692.985707

Joseph K. Goodman, Cynthia E. Cryder, and Amar Cheema. 2013. Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *J. Behav. Dec. Making* 26, 3 (2013), 213–224. DOI: http://dx.doi.org/10.1002/bdm.1753

Sandy J.J. Gould, Duncan P. Brumby, and Anna L. Cox. 2013. What does it mean for an interruption to be relevant? An investigation of relevance as a memory effect. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 57, 1 (September 2013), 149–153. DOI: http://dx.doi.org/10.1177/1541931213571034

Sandy J.J. Gould, Duncan P. Brumby, Anna L. Cox, Víctor González, Dario Salvucci, and Niels Taatgen. 2012. Multitasking and interruptions: a SIG on bridging the gap between research on the micro and macro worlds. In

CHI '12 Extended Abstracts on Human Factors in Computing Systems. CHI EA '12. New York, NY, USA: ACM, 1189–1192. DOI: http://dx.doi.org/10.1145/2212776.2212420

Sandy J.J. Gould, Anna L. Cox, and Duncan P. Brumby. 2013. Frequency and Duration of Self-initiated Task-switching in an Online Investigation of Interrupted Performance. In *Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts AAAI Technical Report CR-13-01*. AAAI, 22–23.

Sandy J.J. Gould, Anna L. Cox, and Duncan P. Brumby. 2015. Task Lockouts Induce Crowdworkers to Switch to Other Activities. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. CHI EA '15. New York, NY, USA: ACM, 1785–1790. DOI: http://dx.doi.org/10.1145/2702613.2732709

Sandy J.J. Gould, Anna L. Cox, Duncan P. Brumby, and Alice Wickersham. 2016. Now Check Your Input: Brief Task Lockouts Encourage Checking, Longer Lockouts Encourage Task Switching. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. DOI: http://dx.doi.org/http://dx.doi.org/10.1145/2858036.2858067

Sandy J.J. Gould, Anna L. Cox, Duncan P. Brumby, and Sarah Wiseman. 2015. Home is Where the Lab is: A Comparison of Online and Lab Data From a Time-sensitive Study of Interruption. *Human Computation* 2, 1 (August 2015), 45–67. DOI: http://dx.doi.org/10.15346/hc.v2i1.4

Tobias Grundgeiger, Penelope Sanderson, Hamish G. MacDougall, and Balasubramanian Venkatesh. 2010. Interruption Management in the Intensive Care Unit: Predicting Resumption Times and Assessing Distributed Support. *Journal of Experimental Psychology: Applied* 16, 4 (December 2010), 317–334. DOI: http://dx.doi.org/10.1037/a0021912

Rikard Harr and Victor Kaptelinin. 2007. Unpacking the social dimension of external interruptions. In *Proceedings of the 2007 international ACM conference on Supporting group work*. GROUP '07. New York, NY, USA: ACM, 399–408. DOI: http://dx.doi.org/10.1145/1316624.1316686

Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '10. New York, NY, USA: ACM, 203–212. DOI: http://dx.doi.org/10.1145/1753326.1753357

Helen M. Hodgetts and Dylan M. Jones. 2006. Interruption of the Tower of London task: Support for a goal-activation approach. *Journal of Experimental Psychology: General* 135, 1 (2006), 103–115. DOI: http://dx.doi.org/10.1037/0096-3445.135.1.103

Eric Horvitz, Johnson Apacible, and Muru Subramani. 2005. Balancing Awareness and Interruption: Investigation of Notification Deferral Policies. In Liliana Ardissono, Paul Brna, & Antonija Mitrovic, eds. *User Modeling 2005*. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 433–437. DOI: http://dx.doi.org/10.1007/11527886_59

Shamsi T. Iqbal and Brian P. Bailey. 2008. Effects of intelligent notification management on users and their tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '08. New York, NY, USA: ACM, 93–102. DOI: http://dx.doi.org/10.1145/1357054.1357070

Shamsi T. Iqbal and Brian P. Bailey. 2005. Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '05. New York, NY, USA: ACM, 1489–1492. DOI: http://dx.doi.org/10.1145/1056808.1056948

Shamsi T. Iqbal and Brian P. Bailey. 2007. Understanding and developing models for detecting and differentiating breakpoints during interactive tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07. New York, NY, USA: ACM, 697–706. DOI: http://dx.doi.org/10.1145/1240624.1240732

Shamsi T. Iqbal and Eric Horvitz. 2007. Disruption and recovery of computing tasks: field study, analysis, and directions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07. New York, NY, USA: ACM, 677–686. DOI: http://dx.doi.org/10.1145/1240624.1240730

Christian P. Janssen, Duncan P. Brumby, and Rae Garnett. 2012. Natural Break Points: The Influence of Priorities and Cognitive and Motor Cues on Dual-Task Interleaving. *Journal of Cognitive Engineering and Decision Making* 6, 1 (March 2012), 5–29. DOI: http://dx.doi.org/10.1177/1555343411432339

Christian P. Janssen, Sandy J.J. Gould, Simon Y.W. Li, Duncan P. Brumby, and Anna L. Cox. 2015. Integrating knowledge of multitasking and interruptions across different perspectives and research methods. *International Journal of Human-Computer Studies* 79 (July 2015), 1–5. DOI: http://dx.doi.org/10.1016/j.ijhcs.2015.03.002

Jing Jin and Laura A. Dabbish. 2009. Self-interruption on the computer: a typology of discretionary task interleaving. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. New York, NY, USA: ACM, 1799–1808. DOI: http://dx.doi.org/10.1145/1518701.1518979

Adam Kapelner and Dana Chandler. 2010. Preventing Satisficing in online surveys. In *CrowdConf*.

Ioanna Katidioti and Niels A. Taatgen. 2014. Choice in Multitasking How Delays in the Primary Task Turn a Rational Into an Irrational Multitasker. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 56, 4 (June 2014), 728–736. DOI: http://dx.doi.org/10.1177/0018720813504216

Nina Keith and K. Anders. 2007. A deliberate practice account of typing proficiency in everyday typists. *Journal of Experimental Psychology: Applied* 13, 3 (2007), 135–145. DOI: http://dx.doi.org/10.1037/1076-898X.13.3.135

Aniket Kittur et al. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. CSCW '13. New York, NY, USA: ACM, 1301–1318. DOI: http://dx.doi.org/10.1145/2441776.2441923

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '08. New York, NY, USA: ACM, 453–456. DOI: http://dx.doi.org/10.1145/1357054.1357127

Steven Komarov, Katharina Reinecke, and Krzysztof Z. Gajos. 2013. Crowdsourcing Performance Evaluations of User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13. New York, NY, USA: ACM, 207–216. DOI: http://dx.doi.org/10.1145/2470654.2470684

Walter S. Lasecki and Jeffrey P. Bigham. 2012. Online quality control for real-time crowd captioning. In *Proceedings of*

the 14th international ACM SIGACCESS conference on Computers and accessibility. ASSETS '12. New York, NY, USA: ACM, 143–150. DOI: http://dx.doi.org/10.1145/2384916.2384942

Walter S. Lasecki, Jeffrey M. Rzeszotarski, Adam Marcus, and Jeffrey P. Bigham. 2015. The Effects of Sequence and Delay on Crowd Work. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. New York, NY, USA: ACM, 1375–1378. DOI: http://dx.doi.org/10.1145/2702123.2702594

Simon Y.W. Li, Ann Blandford, Paul Cairns, and Richard M. Young. 2008. The Effect of Interruptions on Postcompletion and Other Procedural Errors: An Account Based on the Activation-Based Goal Memory Model. *Journal of Experimental Psychology: Applied* 14, 4 (December 2008), 314–328. DOI: http://dx.doi.org/10.1037/a0014397

Simon Y.W. Li, Anna L. Cox, Ann Blandford, Paul Cairns, and A. Abeles. 2006. Further investigations into post-completion error: the effects of interruption position and duration. In *Proceedings of the 28th Annual Meeting of the Cognitive Science Conference*. Vancouver, BC, Canada: Cognitive Science Society, 471–476.

Yikun Liu, Yuan Jia, Wei Pan, and Mark S. Pfaff. 2014. Supporting Task Resumption Using Visual Feedback. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work – Social Computing*. CSCW '14. New York, NY, USA: ACM, 767–777. DOI: http://dx.doi.org/10.1145/2531602.2531710

L.D. Loukopoulos, R.K. Dismukes, and I. Barshi. 2001. Cockpit interruptions and distractions: A line observation study. In *Proceedings of the 11th International Symposium on Aviation Psychology*.

Andrew Mao, Ece Kamar, and Eric Horvitz. 2013. Why Stop Now? Predicting Worker Engagement in Online Crowdsourcing. In *First AAAI Conference on Human Computation and Crowdsourcing*. AAAI, 103–111.

Ke Mao, Ye Yang, Mingshu Li, and Mark Harman. 2013. Pricing Crowdsourcing-based Software Development Tasks. In *Proceedings of the 2013 International Conference on Software Engineering*. ICSE '13. Piscataway, NJ, USA: IEEE Press, 1205–1208.

Gloria Mark, Victor M. González, and Justin Harris. 2005. No task left behind?: examining the nature of fragmented work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '05. New York, NY, USA: ACM, 321–330. DOI: http://dx.doi.org/10.1145/1054972.1055017

Gloria Mark, Shamsi T. Iqbal, Mary Czerwinski, and Paul Johns. 2014. Bored Mondays and Focused Afternoons: The Rhythm of Attention and Online Activity in the Workplace. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*. CHI '14. New York, NY, USA: ACM, 3025–3034. DOI: http://dx.doi.org/10.1145/2556288.2557204

Gloria Mark, Stephen Voida, and Armand Cardello. 2012. "A pace not dictated by electrons": an empirical study of work without email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. New York, NY, USA: ACM, 555–564. DOI: http://dx.doi.org/10.1145/2207676.2207754

Gloria Mark, Yiran Wang, and Melissa Niiya. 2014. Stress and Multitasking in Everyday College Life: An Empirical Study of Online Activity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '14. New York, NY, USA: ACM, 41–50. DOI: http://dx.doi.org/10.1145/2556288.2557361

Winter Mason and Duncan J. Watts. 2010. Financial incentives and the "performance of crowds." *SIGKDD Explor. Newsl.* 11, 2 (May 2010), 100–108. DOI: http://dx.doi.org/10.1145/1809400.1809422

Christopher A. Monk. 2004. The Effect of Frequent Versus Infrequent Interruptions on Primary Task Resumption. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 48, 3 (September 2004), 295–299. DOI: http://dx.doi.org/10.1177/154193120404800304

Christopher A. Monk, Deborah A. Boehm-Davis, George Mason, and J. Gregory Trafton. 2004. Recovering From Interruptions: Implications for Driver Distraction Research. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 4 (Winter 2004), 650–663. DOI: http://dx.doi.org/10.1518/hfes.46.4.650.56816

Christopher A. Monk, J. Gregory Trafton, and Deborah A. Boehm-Davis. 2008. The Effect of Interruption Duration and Demand on Resuming Suspended Goals. *Journal of Experimental Psychology: Applied* 14, 4 (December 2008), 299–313. DOI: http://dx.doi.org/10.1037/a0014402

Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgm Decis Mak* 5, 5 (August 2010), 411–419.

Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behav Res* 46, 4 (December 2014), 1023–1031. DOI: http://dx.doi.org/10.3758/s13428-013-0434-y

Raj M. Ratwani, Alyssa E. Andrews, Jenny D. Sousk, and J. Gregory Trafton. 2008. The Effect of Interruption Modality on Primary Task Resumption. In *Human Factors and Ergonomics Society Annual Meeting Proceedings*. Sage, 393–397. DOI: http://dx.doi.org/10.1177/154193120805200441

Jeffrey M. Rzeszotarski, Ed Chi, Praveen Paritosh, and Peng Dai. 2013. Inserting Micro-Breaks into Crowdsourcing Workflows. In *Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts AAAI Technical Report CR-13-01*. AAAI, 62–63.

Jeffrey M. Rzeszotarski and Aniket Kittur. 2012. CrowdScape: interactively visualizing user behavior and output. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. UIST '12. New York, NY, USA: ACM, 55–62. DOI: http://dx.doi.org/10.1145/2380116.2380125

Jeffrey M. Rzeszotarski and Aniket Kittur. 2011. 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 '11. New York, NY, USA: ACM, 13–22. DOI: http://dx.doi.org/10.1145/2047196.2047199

Dario D. Salvucci. 2010. On reconstruction of task context after interruption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '10. New York, NY, USA: ACM, 89–92. DOI: http://dx.doi.org/10.1145/1753326.1753341

Dario D. Salvucci and Peter Bogunovich. 2010. Multitasking and monotasking: The effects of mental workload on deferred task interruptions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '10. New York, NY, USA: ACM, 85–88. DOI: http://dx.doi.org/10.1145/1753326.1753340

Dario D. Salvucci, Niels A. Taatgen, and Jelmer P. Borst. 2009. Toward a unified theory of the multitasking continuum:

from concurrent performance to task switching, interruption, and resumption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. New York, NY, USA: ACM, 1819–1828. DOI: http://dx.doi.org/10.1145/1518701.1518981

Harini Alagarai Sampath, Rajeev Rajeshuni, and Bipin Indurkhya. 2014. Cognitively Inspired Task Design to Improve User Performance on Crowdsourcing Platforms. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*. CHI '14. New York, NY, USA: ACM, 3665–3674. DOI: http://dx.doi.org/10.1145/2556288.2557155

Ognjen Scekic, Christoph Dorn, and Schahram Dustdar. 2013. Simulation-Based Modeling and Evaluation of Incentive Schemes in Crowdsourcing Environments. In Robert Meersman et al., eds. *On the Move to Meaningful Internet Systems: OTM 2013 Conferences*. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 167–184. DOI: http://dx.doi.org/10.1007/978-3-642-41030-7_11

Brian Shackel. 1962. Ergonomics in the Design of a Large Digital Computer Console. *Ergonomics* 5, 1 (1962), 229–241. DOI: http://dx.doi.org/10.1080/00140136208930578

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 254–263.

Studs Terkel. 1974. *Working: People Talk About What They Do All Day and How They Feel About What They Do*, New York, NY, USA: Random House.

J. Gregory Trafton, Erik M. Altmann, and Raj M. Ratwani. 2011. A memory for goals model of sequence errors. *Cogn Syst Res* 12, 2 (June 2011), 134–143. DOI: http://dx.doi.org/16/j.cogsys.2010.07.010

Roelof Anne Jelle de Vries, Manja Lohse, Andi Winterboer, Frans C.A. Groen, and Vanessa Evers. 2013. Combining social strategies and workload: a new design to reduce the negative effects of task interruptions. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '13. New York, NY, USA: ACM, 175–180. DOI: http://dx.doi.org/10.1145/2468356.2468388

E. Walther, R. Weil, and J. Dusing. 2011. The Role of Evaluative Conditioning in Attitude Formation. *Current Directions in Psychological Science* 20, 3 (June 2011), 192–196. DOI: http://dx.doi.org/10.1177/0963721411408771

Nicole E. Werner, David M. Cades, Deborah A. Boehm-Davis, Jessica Chang, Hibah Khan, and Gia Thi. 2011. What Makes Us Resilient to Interruptions? Understanding the Role of Individual Differences in Resumption. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 55, 1 (September 2011), 296–300. DOI: http://dx.doi.org/10.1177/1071181311551062

Johanna I. Westbrook et al. 2010. The impact of interruptions on clinical task completion. *Qual Saf Health Care* 19, 4 (2010), 284–289. DOI: http://dx.doi.org/10.1136/qshc.2009.039255