Human Computation (2015) 2:1:45-67
© 2015, Gould et al. CC-BY-3.0
ISSN: 2330-8001, DOI: 10.15346/hc.v2i1.4

Home is Where the Lab is: A Comparison of Online and Lab Data From a Time-sensitive Study of Interruption

SANDY J. J. GOULD, UCL INTERACTION CENTRE, UNIVERSITY COLLEGE LONDON

ANNA L. COX, UCL INTERACTION CENTRE, UNIVERSITY COLLEGE LONDON

DUNCAN P. BRUMBY, UCL INTERACTION CENTRE, UNIVERSITY COLLEGE LONDON

SARAH WISEMAN, UCL INTERACTION CENTRE, UNIVERSITY COLLEGE LONDON

ABSTRACT

While experiments have been run online for some time with positive results, there are still outstanding questions about the kinds of tasks that can be successfully deployed to remotely situated online participants. Some short tasks, such as menu selection, have produced viable results but these do not represent the gamut of tasks that interest human-computer interaction researchers. In particular, we wondered whether long-lasting, time-sensitive tasks that require continuous concentration could work successfully online, given the confounding effects that might accompany the online deployment of such a task. We ran an archetypal interruption experiment online and in the lab to investigate whether studies that are long-lasting and time-sensitive might be more vulnerable to a loss of control than the short, time-insensitive studies that are representative of the majority of previous online studies. Statistical comparisons showed no evidence of performance differences across a number of dimensions. Despite these results, there were issues with data quality that stemmed from participants misunderstanding the task. Our findings suggest that long-lasting experiments using time-sensitive performance measures can be run online but that care must be taken when introducing participants to experimental procedures.

1. INTRODUCTION

There is a long tradition in human-computer interaction (HCI) of running controlled lab studies (see, e.g., Newell & Card, 1985). Running a study in a lab is resource intensive. As many experimental tasks are performed on computers, researchers have decided to administer studies using the Internet instead of having participants come to a lab (e.g., Snow et al., 2008; Suri &

Watts, 2011). Although there are obvious concerns over diminished experimental control when experiments are run online, systematic efforts to compare data suggest that some kinds of HCI research can reliably be conducted with remote online participants (Dandurand, Shultz, & Onishi, 2008; Heer & Bostock, 2010; Komarov, Reinecke, & Gajos, 2013). However, these studies have tended to focus on simple tasks that can be completed relatively quickly. For example, online studies have been useful in determining which menu layouts allow users to make selections most quickly (Komarov et al., 2013). While such studies have certainly made useful contributions to their respective fields, many tasks that HCI researchers are interested in are neither simple nor quick. It is less clear whether longer-lasting tasks can be investigated with online participants and still produce high quality data.

Routine procedural tasks represent an important class of activity across a number of domains. Well-practiced procedures are employed whenever pilots program co-ordinates into flight computers, nurses program parameters into infusion pumps or commuters purchase railway tickets from automated kiosks. These tasks all require users to program an interface using an explicit procedure that can take anywhere from a few seconds to a few minutes to perform. Researchers have investigated how people perform routine procedural tasks in a variety of scenarios, including how multiple tasks are interleaved. There has been significant attention devoted to studying the effects of interruption on performance (e.g., Cades, Davis, Trafton, & Monk, 2007; Chung & Byrne, 2008; Monk, Trafton, & Boehm-Davis, 2008). Interruption studies typically ask participants to execute a series of tasks in each trial while dealing with occasional interruptions. Usually, these interruptions are randomly distributed, so participants have to attend to the task continuously or risk making costly errors. In these paradigms, any distractions arising during online participation are likely to have a negative effect on performance. Unanticipated performance deviations are exacerbated by the structure of such experiments; accommodating interruptions at random intervals means increasing trial time and so reducing the number of trials that can be run. In-turn this amplifies the effects of noisy trials on a sample. Thus, there are good reasons to think that studies of routine procedural action might work poorly online.

Here we investigate whether studies that require participants to work on time-sensitive routine procedural tasks can be administered remotely. We focus on the effect of interruptions on performance and compare performance data from participants that worked online with those who participated in the lab. The task we used was long-lasting and required almost continuous attention from participants. We anticipated that these characteristics would result in poor performance and low-quality data when people took part sitting in their living rooms on their laptops. In practice we found that there was no statistically significant difference between online and lab data over a number of performance measures. This paper contributes a rigorous comparison of lab and online data in the context of a relatively long-lasting, time-sensitive experiment that requires continuous concentration.

1.1. Crowdsourcing

Crowdsourcing is a method for marshalling distributed sets of individuals to achieve a common goal. On paid crowdsourcing platforms individuals participate to earn money (e.g., Araujo, 2013; Chandler, Mueller, & Paolacci, 2014; Shaw, Horton, & Chen, 2011). There is also a tradition of volunteer crowds contributing to citizen science projects (e.g., Jennett et al., 2014; Lintott et al., 2008; Moore et al., 2011).

Crowdsourcing is often sold on the basis that it is fast, cheap and provides access to large populations. The size of the Amazon Mechanical Turk (AMT) population is estimated to be in excess of 500,000 workers (Paolacci & Chandler, 2014), many of whom are available at any given time. Diverse groups can thus be recruited very quickly. Low rates of pay have often been seen as a positive aspect of crowdsourcing (e.g., Buhrmester, Kwang, & Gosling, 2011; Snow et al., 2008), although there is increasing recognition of the ethical issues related to low pay (see, e.g., Kittur et al., 2013). Regardless of remuneration levels, crowdsourcing remains low cost in the sense that the time and effort that researchers and requesters must invest in organizing work is substantially reduced compared with running laboratory-based studies.

Research with crowds can be broadly thought of as *using* crowds or *investigating* crowds. Crowds can be used to investigate human behaviour. For example, judgements about risk-taking (Eriksson & Simpson, 2010), the emotional effects of multitasking (B. Morgan & D'Mello, 2013) and the functioning of prospective memory (Gilbert, 2015) have all been investigated using AMT. In these studies crowdsourcing platforms are used in the same way as participant pools at university psychology departments; the focus is on trying to understand a particular phenomenon and the sampling technique is largely incidental.

A different strand of research has viewed crowds themselves as a phenomenon to be understood. For example, researchers have explored the demographic characteristics of crowdsourcing platforms (Paolacci & Chandler, 2014; Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010), or ways to optimally price different assignments (Hsieh, Kraut, & Hudson, 2010; Mason & Watts, 2010; Shaw et al., 2011). Other work has taken crowdsourcing beyond single worker microtasks. Crowds assembled in real-time to collaborate on activities are both feasible and useful (Bernstein, Brandt, Miller, & Karger, 2011; Bernstein, Karger, Miller, & Brandt, 2012; Lasecki & Bigham, 2012).

In this paper we contribute to both of the research traditions just described: we study a phenomenon using participants from a crowdsourcing platform while at the same time developing our understanding of the particular characteristics of data collected online. We are concerned with how differences between crowds and traditional populations (i.e., crowds as a phenomenon) can affect the results of studies that seek to understand human behaviour. We situate this work in the

space of comparative online-offline studies (e.g., Dandurand et al., 2008; Germine et al., 2012; Komarov et al., 2013).

1.2. Online data collection

Some of the first online data collection focused on obtaining survey data (Stanton, 1998) and it is now routine to deploy survey studies online because online collection is seen as an inexpensive yet reliable source of data (Behrend, Sharek, Meade, & Wiebe, 2011). More recently, there has been a growing contingent of researchers investigating human performance through experiments delivered online, whether to better understand what normal performance looks like for traditional crowdsourced work (Rzeszotarski & Kittur, 2011), or to perform novel research (Suri & Watts, 2011).

Although online experimentation has been successful in many cases, researchers have remained mindful of the loss of control that is an inevitable consequence of moving experiments online. A number of studies have been conducted that compare the quality of data collected online with data collected through traditional lab procedures.

Dandurand et al. (2008) compared online and lab-collected data from a problem-solving experiment. Although online participants were less accurate in their study, the results were considered to be practically equivalent to those gained through lab experimentation. Paolacci et al. (2010) conducted a similar comparative study, also using problem-solving tasks. They used the AMT crowdsourcing platform to recruit participants, rather than the subject pool employed by Dandurand et al. Opening a study to the Internet at large is more risky than using a well-known subject pool, but Paolacci et al. still came to the same conclusion: data collected online were equivalent to those collected in the lab. This equivalency has also been demonstrated for survey-based investigations (Buhrmester et al., 2011) and other experimental paradigms (e.g., Germine et al., 2012; Heer & Bostock, 2010; Mason & Suri, 2012).

Despite the success of these studies in demonstrating equivalence between lab and online data, the generalizability of their conclusions is constrained by their experimental paradigms. In particular, surveys and problem-solving studies are likely to be less sensitive to any loss of control during an online experimentation than, for example, studies in which time is a critical factor being measured. In time-sensitive experiments, uncontrollable participant behaviour, such as interleaving participation with other tasks, could confound timing data. Where the outcome measure is categorical, interleaving presents less of a problem than in experiments that use reaction times or other time-sensitive measures of performance. Table 1 lists the number and duration of trials in four comparative studies of data collected online and in the lab. Of the studies listed in Table 1, only one has a time-sensitive primary measure (Komarov et al., 2013), although duration and other timing data is recorded in the other three studies.

Work by Komarov, Reinecke, and Gajos (2013) reported a comparison of three time-sensitive user interface experiments that were run simultaneously in the lab and online. Komarov et al. found no significant differences in the data from the two sources. This finding has important implications for researchers investigating routine procedural tasks because it suggests that time-sensitive experiments work just as well online as other kinds of experiments.

While Komarov et al.'s study provides further strong evidence for the reliability of online experimentation, like previous studies, its generalizability is in some ways still limited. Specifically, it is not clear whether their results would transfer to other domains where tasks are not necessarily as short (i.e., under 2 seconds per trial) or atomizable as they were in Komarov et al.'s work. Repetitive, routine tasks in the crowdsourcing literature are typically split into very short micro-tasks; for more skilled work, multi-hour units of work might be assigned (Kittur et al., 2013). Crowdworkers might typically work for four hours per day on average (Lasecki, Rzeszotarski, Marcus, & Bigham, 2015), individual tasks are usually very much shorter. Periods where undivided attention is required for up to an hour are atypical in paid crowdsourcing settings. The evidence collated in Table 1 reinforces the idea that concentration is typically required in short bursts in online studies. Where trials are longer, as in Dandurand et al. (2008), there are typically much fewer of them.

Study	Trial duration (s)	Trials
Komarov et al. (2013)	1.5	360
Dandurand et al. (2008)	134	16
Heer and Bostock (2010)	40	62
Kittur et al. (2008)	90	4

Table 1. Overview of total number and duration of trials in selected comparative studies

1.3. Interruptions and multitasking

Unfortunately, decomposing tasks is not always compatible with investigations of routine procedural tasks, or with interruption studies in particular, because such studies are necessarily constrained in ways that make them difficult to decompose into smaller tasks. Interruption studies often have small numbers of trials because each trial needs to be long enough so that there are substantial periods of work on a primary task between interruptions. As a consequence, interruption experiments often have fewer, longer trials yielding fewer measurements than in other kinds of experiments.

Having fewer trials means there is greater potential for noise to distort results (for instance, caused by participants going to make themselves coffee or responding to instant messages during

the task), so the loss of control entailed by online experimentation has the potential to have a larger effect on results than it might have in an experiment with a large number of short trials. For example, one of Komarov et al.'s (2013) experiments had participants perform 360 trials of approximately 1.5 seconds each. Such an approach would not be possible with an interruption study. Other work has had longer trial times, although in the case of Heer and Bostock (2010), online trial times were four times longer than were expected from lab observations, reinforcing concerns over deploying time-sensitive studies online. Other work has utilized problem-solving or non-routine tasks (Dandurand et al., 2008), which can be problematic when used in interruption studies (Salvucci, 2010) and whose results are not applicable to the kinds of routine procedural tasks that we focus on.

Another feature of typical interruption studies, as well as a raft of psychology-influenced HCI studies, is that they often investigate memory. Altmann and Trafton's Memory for Goals theory (Altmann & Trafton, 2002) is an account of the memory processes involved in goal suspension, rehearsal and recovery during interruptions. A large number of experiments drawing on the theory have been reported (e.g., Altmann, Trafton, & Hambrick, 2013; Li, Blandford, Cairns, & Young, 2008; Monk et al., 2008; Ratwani & Trafton, 2008; Trafton, Altmann, & Ratwani, 2011), all of which require participants to rely on memory alone to keep track of their progress throughout trials. A potential issue with conducting these kinds of experiments online is that should a participant switch away from the experiment to do another task, there is the potential for unprompted forgetting behaviour to have a sizeable effect on performance. This is because the tasks typically used in these experiments provide no progress cues to participants that could aid resumption. Such participant-generated interruptions and other discretionary multitasking behaviour, could obscure the effect of experimentally generated interruptions.

Discretionary multitasking and interleaving of tasks have been studied in laboratories (Payne, Duggan, & Neth, 2007; Salvucci & Bogunovich, 2010) and self-interrupting behaviour has been studied in workplaces (Dabbish, Mark, & González, 2011; González & Mark, 2004). These studies have shown that people are frequently interrupted (González & Mark, 2004; Mark, Gonzalez, & Harris, 2005) and that interruptions can increase stress (Mark, Gudith, & Klocke, 2008). They also show that multitasking strategies are nuanced and contingent on the constraints of the task (Salvucci, 2010; Salvucci & Bogunovich, 2010) and the context in which it is performed (Dabbish et al., 2011). The role that tasks play in shaping multitasking behaviour means that care needs to be taken when comparing studies of multitasking conducted in different environments and using different methods. If we want to be able to study interruptions in online settings, we first need to understand how online environments affect behaviour relative to a well-understood baseline.

This paper presents a comparison of online and lab-collected data from an experiment that has features that could make it acutely sensitive to the negative effects of online deployment. The experiment uses a routine procedural data-entry task that has been designed to examine the effect

of predetermined interruptions on place-keeping performance in a routine task (e.g., Gray, 2000). In the lab there is minimal scope for participants to create additional interruptions through self-interruption and task switching. However, when completing an experiment online – with no experimenter peering over their shoulder – participants might find any number of unprompted distractions (Dabbish et al., 2011) during the experiment, from instant messaging to tea-making. Both device-generated and participant-generated interruptions will have disruptive effects just like any interruption, except that unlike the interruptions generated by the experiment itself, they are likely to affect performance in ways that cannot be anticipated.

2. METHOD

1.1. Participants

Forty-eight participants took part in the study. Twenty-four participants (15 female) with a mean age of 24 years (*SD*=6 years) took part in the lab study. Twenty-four participants (13 female) with a mean age of 29 years (*SD*=9 years) took part in the online study.

All participants were drawn from the same university subject pool and were paid £7 (~$11) for approximately one hour of their time. Both lab and online participants signed-up for the study through an online subject pool system. The process of online participation was automated; the experimenters' only role was to approve payments.

Although it has been shown that crowdsourcing platforms such as MTurk produce similar data to that produced in lab studies, we used a subject pool because it removed the potential for issues stemming from underlying differences in the populations sampled. Additionally participants in the subject pool were likely to have experience of taking part in experiments, reducing any effects stemming from use of a different population that might be less familiar with participating in lab-style experiments that use routine procedural tasks. Lab participants were paid in cash, online participants were paid in Amazon vouchers. None of the participants had experience with the task before starting the experiment.

1.2. Design

The experiment used a 2x2x2 mixed design. There were two within-subjects independent variables: interruption relevance, which had two levels, relevant and irrelevant; and interruption timing, which had two levels, within-subtask and between-subtask. There was one between-subjects independent variable, experiment location, which had two levels, online and lab. The primary dependent variable was *resumption lag* (Altmann & Trafton, 2002), the time it took participants to resume working on the primary task after being interrupted. We also collected other quantitative measures of performance such as error rate and trial duration.

Our primary focus in this study was context – the extent to which the setting in which participants took part (online or in the lab) affected their performance in a number of measures that are standard in interruption studies.

1.3. Materials

The task used in this experiment was the *Pharmacy Task*, an adaptation of the *Doughnut Machine* (Li, Cox, Blandford, Cairns, & Abeles, 2006). The Pharmacy Task is a routine data entry task that has been used previously to investigate the effects of interruptions (Gould, Brumby, & Cox, 2013). Participants were given a set of '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, Colour, Packaging, Label* (Figure 1). Participants completed one subtask at a time. After entering the values for a particular subtask, they clicked the 'OK' button at the bottom of each subtask. The entered values were cleared and participants then started working on the next subtask.

From time-to-time, participants were interrupted. There were two interruptions per trial that arrived either between one subtask finishing and another starting (e.g., *Packaging* had just been finished, but *Label* had not been started) or in the middle of working on a subtask (i.e., at least one value had already been entered, but the subtask had not been completed). The content of the interruption varied with each condition. All interruptions comprised two audit tasks interposed between three transcription tasks. The transcription task was a simple filler task that required participants to copy four-digit values from a list into corresponding text fields. The transcription task was the same irrespective of condition.

Unlike transcription tasks, the content of audit tasks varied with each condition. Audit tasks asked participants a question about either their progress through the task (relevant) or about their knowledge of the various components of the task (irrelevant). After all five components of an interruption were completed, participants were returned to the primary task, where they attempted to resume where they were about to work before they were interrupted. This task is made more difficult because any cues that might aid resumption, such as values that have already been entered, are hidden from the interface at the moment of resumption. Participants had to resume precisely where they left off before they were interrupted; if they were about to work on the *Red* element of the *Colour* subtask when they were interrupted, they would need to resume on the *Red* element of the *Colour* subtask. If participants selected the wrong subtask, they were given an eight-second lock-out penalty. During this time the components of the task were hidden and participants could not interact with the task. If participants selected the correct subtask but the wrong element (e.g., they selected *Capsule* when they should have selected *Lozenge*), they lost any progress that had been made on the subtask, and had to start the subtask from scratch. These penalties were introduced to discourage guessing behaviour on resumption.

After completing the experiment, participants were given a twelve-item questionnaire, which asked them about their experience of the task and the interruptions. Participants indicated their agreement with the statements on a five-point Likert scale. In addition to the task and questionnaire, the materials also included written instructions, an introductory video and a debriefing. Online participants had additional information related to the mechanics of online participation.

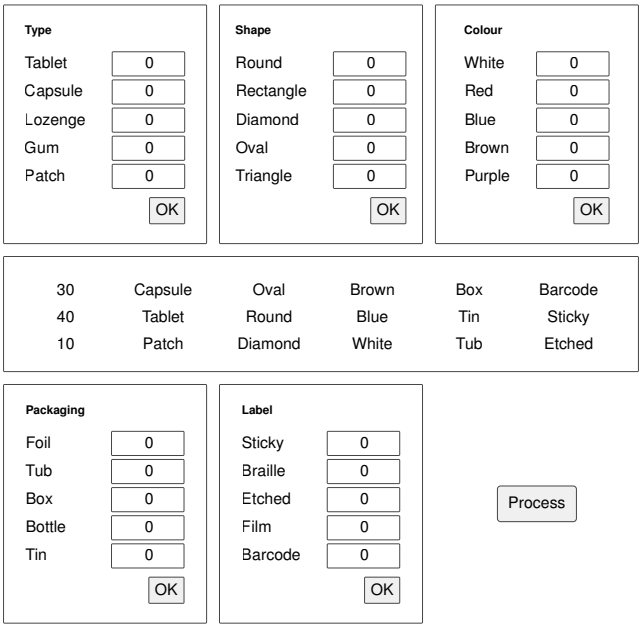

Figure 1. ***The Pharmacy Task, as rendered in the browser. Subtasks are completed left-to-right, top-to-bottom***

1.4. Implementation

For both conditions the experiment was implemented in the browser using HTML5 techniques including audio and video. Browser compatibility was confirmed with WebKit browsers, Firefox and Internet Explorer 9+.

To reduce the possibility of transient network outages breaking the study, the whole study was loaded into a single page and the components were selectively shown to the participants as they progressed through the study. To prevent accidental navigation away from the experiment, participants were prompted to confirm their desire to leave with a modal dialog. However, there was no way to resume the experiment if participants closed the page or in the event of system or browser failure.

1.5. Procedure

1.1.1. Online

Participants signed-up through a section in a university subject pool specifically for online studies. Once they had agreed to participate they were given a link that was valid until the end of their timeslot; participants could take part in the experiment at any moment up to the end of their slot, but not after it had expired. The experimenters' only involvement in this process was to pay participants after they had completed the study.

Clicking the link took participants from the subject pool to the website hosting the experiment. On the first screen participants were given information about the online aspects of the experiment and a widget to test whether their sound was working. Participants were told that this was an experiment concerned with how people execute routine tasks and that it was important that they were not disturbed by anything or anyone during the study. Participants were told that their compensation depended on not having any long breaks during the study, although in reality all participants who finished the experiment were paid the same amount. On the next screen the experiment was described to participants and they were given an instructional video to watch. After finishing with the introductory materials, participants moved on to the training phase of the experiment. There were a total of four training trials; the first was uninterrupted and the other three had interruptions drawn from different conditions. Once participants had completed the training trials they completed twelve experimental trials. There were breaks of up to one minute after four and eight trials, which participants could skip if they wanted.

Once participants had completed the post-experiment questionnaire and read the debriefing they were returned to the subject pool where their participation was automatically registered. They were then emailed an Amazon gift voucher.

1.1.2. Lab

The procedure for participants taking part in the lab was kept as close as possible to the one used online. Participants signed up for the experiment through the same university subject pool. They were given the same instructions (except for the instructions concerned specifically with the online nature of the study) and watched the same introductory video. Participants completed the same number of training and experimental trials. The biggest difference between the online and lab procedures was that participants in the lab did the training trials with the assistance of the experimenter. This afforded participants the opportunity to ask questions about the task or the functioning of a particular interface element and allowed the experimenter to correct any obvious deficiencies in participants' knowledge. After the experiment finished, participants completed the same questionnaire as the online group, were verbally debriefed and paid in cash.

3. RESULTS

The goal was to compare the quality of lab-collected and online-collected data sampled from a single population. To this end we undertook two conceptually different kinds of analyses. In the first set of analyses *(Processed data)* we use data cleaning techniques to minimize the impact of outliers on our results. These techniques are standard both in laboratory investigations of interruptions (e.g., Brumby, Cox, Back, & Gould, 2013) and in online experiments. Indeed, in the crowdsourcing domain, tools have been developed specifically for aiding this process (e.g., Rzeszotarski & Kittur, 2012). Such techniques allow us to make conclusions based on the most likely behaviour arising in a given scenario; the influence of a few extreme edge cases to radically shift the mean is minimized.

The prevalence and nature of edge cases is not a concern for most studies, but we were mindful that the purpose of our investigation was to understand the effects of different approaches to data collection on results obtained. Outliers and poor performance might provide a source of insight into potential differences in results between online and lab-based methods of data collection. These results are reported in the *Outliers and Raw Data* subsection.

1.1. Processed data

Online and lab data were processed through identical steps. The first step was to discard any participants who had fewer than two correct post-interruption resumptions out of the six interruptions that participants encountered for a given condition. These participants were removed because making so few correct resumptions suggested that participants had struggled to understand the instructions.

A total of five participants in the online condition produced no correct resumptions for one or more conditions and were discarded from the sample. One participant regularly took over 20 seconds to resume after an interruption, suggesting that their attention was elsewhere during the course of the experiment. Their data were also discarded. This meant that a total of six participants' data were discarded from the online group. Under the same criteria, no participants needed to be discarded from the lab group. This left uneven group sizes for the comparison and some caution was therefore taken in the calculation of statistics; we used appropriate sums of squares and non-parametric tests where appropriate and possible.

The second step was to remove any task resumptions that were incorrect. A resumption was considered incorrect if participants selected any subtask or subtask element other than the one they were supposed to return to on resumption. We removed these incorrect resumptions because strategic adaptations associated with recovering after errors, such as backtracking, can skew reaction time data. Fourteen per cent of resumptions in the lab data (83/576) were incorrect, while 16% of resumptions were incorrect in the lab condition (70/431). We report the distribution of these errors later.

The third step was to remove outliers on a trial by trial basis. The mean resumption lag for each condition was calculated for each participant individually. Any resumptions greater than 1.96 standard deviations from the mean resumption lag (i.e., 95%, a common threshold in experimental work) were discarded from the sample. We did not consider resumptions outside the threshold to be representative of normal performance. Participants may have been distracted during this period and our focus in this study is on what the normal process of post-interruption resumption looks like. Komarov, Reinecke, and Gajos (2013) have suggested that interquartile ranges may be more useful than the 95% measure in circumstances where there are extreme outliers, but this was unnecessary for our sample. Of the 361 resumptions remaining in the online condition after removing errors, a further 21 (6%) resumptions were removed as outliers. Twenty-one outliers (4%) of 493 correct resumptions were also removed from the lab condition.

1.1.1. Resumption lags

The main dependent variable in this experiment was resumption lag. This is the time it takes to resume working on a task after an interruption has finished. It is commonly used as a measure of the disruptiveness of an interruption because taking longer to resume after an interruption implies having to think harder about what was being done before being interrupted (Hodgetts & Jones, 2006; Trafton et al., 2011). In the absence of a suitable non-parametric test, a mixed factorial ANOVA with unequal group sizes was used to analyse the results.

Participants took around 4000-ms to accurately resume activity on the primary task after interruption. Fastest resumptions occurred in the lab group after relevant, within-subtask interruptions (M=3456-ms, SD=1652-ms), while slowest resumptions also occurred in the lab group but instead after irrelevant, between-subtask interruptions (M=4680-ms, SD=1938-ms). These results are illustrated in Figure 2.

There was no significant main effect of experiment location (online vs lab), $F(1,40)=0.02$, $p=0.88$, or of interruption timing (between- vs within-subtask), $F(1,40)=4.00$, $p=.053$. However, there was a significant main effect of interruption relevance (relevant vs irrelevant) $F(1,40)=9.62$, $p<.01$, $\eta_p^2=.20$.

There were four possible combinations of interaction effects (location × type × timing; location × type; type × timing; location × timing), none of which were significant.

1.1.2. Error rates and audit accuracies

After completing an interruption, participants had to resume the primary task where they left off. However, as the task removed all resumption cues, this was somewhat difficult. Participants sometimes attempted to resume somewhere other than where they had left off: they made resumption errors. Participants in the lab condition made 83 resumption errors out of a total of 576 resumptions (14%). There were fewer participants in the online condition so there were

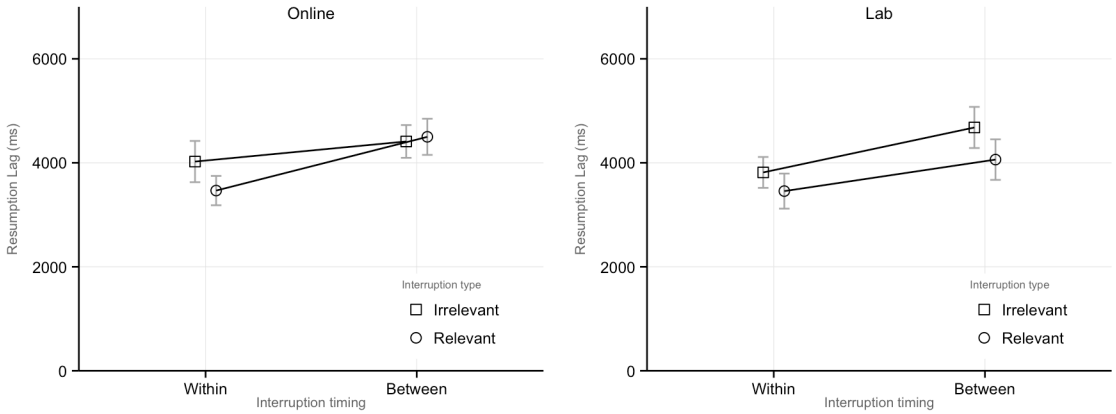

Figure 2. Split interaction plot of resumption lags collected online (left) and in the lab (right). All times in milliseconds

correspondingly fewer resumptions: from a total of 432 resumptions, participants resumed incorrectly on 70 occasions (16%). The number of resumption errors ranged from 0-8 (M=3, SD=3) for lab participants and 1-13 for the online participants (M=4, SD=3). Accounting for the uneven samples, a Mann-Whitney test showed no significant difference in error rates between online and lab participants (U=207, z=0.23, p=0.82). Variation by condition was somewhat higher than on aggregate, but the samples for each condition were very small because many participants made no errors in a particular condition. This allowed a single participant to skew the error rate in any particular condition. They should therefore be interpreted illustratively.

	Within				Between	
	Element		Subtask		Subtask	
	Lab	Online	Lab	Online	Lab	Online
Relevant	8%	16%	3%	5%	13%	16%
Irrelevant	14%	11%	2%	3%	17%	15%

Table 2. Resumption error rates by condition.

The kinds of errors that participants could make depended on the experimental condition. When participants were resuming after a within-subtask interruption, they could choose either the wrong subtask element or they could choose the wrong subtask. When resuming after between-subtask interruptions, there was no particular subtask element that participants needed to work on when they resumed, so there were no 'element' resumption errors occurring after between-subtask interruptions. A breakdown of resumption errors by condition is given in Table 2.
During the interruptions, participants completed both transcription tasks and audit tasks. Transcription tasks required participants to enter the correct values before they could proceed. Interposed between the transcription tasks were audit tasks. Relevant audit tasks required participants to recall something relevant to their current progress through the task. Irrelevant audit tasks required participants to recall some fact that was related to the task at hand but that did not require them to consider their current progress.

To avoid giving participants information that might aid resumption after interruption, no feedback was given to answers to audit questions and consequently no penalty could be levied on guessing behaviour. Participants could choose from one of five answers in each audit task, so guessing would have produced an audit accuracy rate of 20%. The data show that participants' accuracy was far superior to guessing and was broadly similar whether the participant was working online or in the lab. A breakdown of audit accuracy by condition is given in Table 3. Overall, online participants ranged from 11 to 47 audit errors (M=27, SD=12) during the experiment (of a total of 48 opportunities for error). The range for lab participants was 11-43 (M=25, SD=8). There was no significant difference in audit accuracy as a function of mode of participation (U=211, z=0.13, p=0.91).

	Within		Between	
	Lab	Online	Lab	Online
Relevant	61%	54%	47%	50%
Irrelevant	51%	60%	53%	57%

Table 3. Audit accuracy by condition

1.1.3. Trial durations

Timing data were collected at all stages of the experiment but it was felt that the most useful measure of the time it took participants to complete the experiment was the average duration of experimental trials (i.e., excluding training). This was because total experimental time would include the time spent on discretionary rests and could be distorted by anomalously long trials. Each participant completed a total of twelve experimental trials. After applying the same 95% filter to the trial times as was used for resumption lags, the total duration of a trial from start to finish was aggregated across all conditions. Participants who did the experiment in the lab took a

mean of 124-s to complete each trial (*SD*=38-s), whereas participants in the online condition took a mean of 137-s (*SD*=39-s). The results of a Mann-Whitney test showed that this difference was not significant (U=250, z=0.86, p=0.38).

We also wanted to understand the performance profile of the six participants who were excluded from the resumption lag analysis. These participants took an average of 181-s for each trial (SD=47-s). Holm-corrected post hoc Mann-Whitney tests showed that excluded participants were significantly slower to complete each trial than both the online participants included in the analysis (U=88, z=2.27, p<.05) and the lab participants (U=124, z=2.70, p<.01).

1.2. Outliers and raw data

1.1.1. Participants discarded from the online condition

Six participants were discarded from the analysis of online participants' performance. These participants represented 25% of the sample. No participants were discarded from the lab sample. These participants represent a source of variation so it is important to explore why these six participants performed as they did.

Of the six participants who had their data excluded, five made a sufficient number of errors that their data could not meaningfully be analyzed. We focus on the activity of these five participants. The sixth participant took over 18 seconds to resume after interruptions and was deemed to be disengaged from the study. We examined the excluded participants' task execution on a step-by-step basis. The goal of this approach was to establish the cause of the excluded participants' poor performance. Specifically, we wanted to know whether the poor performance was the result of systematic deficiencies in participants' understanding of the task or whether it was simply due to a lack of attentiveness during task execution. This distinction is important; deficiencies in the understanding of the task can be corrected with better pre-experiment materials. Attentiveness issues stemming from online administration are potentially more difficult to address.

Participants each completed a total of twelve trials. There were two interruptions in each trial. Each participant therefore encountered a total of 24 interruptions. This meant that participants resumed after interruptions six times for each of the four conditions. If there were fewer than two correct post-interruption resumptions out of the six for each for each condition, a mean resumption lag could not be meaningfully calculated. If a participant had no usable data for any one of the conditions we discarded all data from that participant. The question is thus whether the five participants were discarded due to a particular class of error, or were generally poor at the task.

We looked closely at the kinds of resumption errors that the five discarded participants made. When participants returned to the task after an interruption they were presented with

decontextualized interface. Both target information and progress were hidden. Participants had to remember what they were doing before the interruption. When resuming participants could make errors by selecting the wrong subtask (e.g., *Type* instead of *Shape*) or the wrong subtask element (e.g., *Tablet* instead of *Patch*). We looked at the relative prevalence of these errors.

The five discarded participants made a total of 63 resumption errors, from a total of 120 resumptions (53%). Of these 63 resumption errors, 21 resumption errors occurred because participants selected the wrong subtask when they resumed. The remainder occurred because participants picked the correct subtask but attempted to work on the wrong subtask element. One participant accounted for nine of the incorrect subtask choices. The other participants were far more likely to choose the correct subtask and then pick the incorrect subtask element.

Thirty of the 63 errors (48%) occurred when participants selected the subtask element they had just completed before the interruption. For instance, they may have entered *20* for *Lozenge*, *30* for *Capsule* and then been interrupted. Interruptions that occurred while a participant was working on a subtask (i.e., within-subtask interruptions) always occurred as participants went to work on the *next* subtask element. Participants were told to continue from exactly the point they left off in the primary task. For these thirty errors, participants went back to the last task they had *completed*. One of the five discarded participants made this error for every one of the twelve resumptions after a within-subtask interruption.

4. DISCUSSION

Our results show that across a number of relevant measures, there were no statistically significant differences in the data produced by participants who took part in the experiment online and those who did the experiment in the lab. This suggests that online interruption research is viable and, more broadly, experiments involving routine procedural tasks that are long in duration, are time-sensitive, require continuous concentration, and generate relatively small quantities of data. Additionally, our results augment previous comparative work in the area that has come to the same conclusions (Dandurand et al., 2008; Komarov et al., 2013; Paolacci et al., 2010), albeit in different domains with different kinds of tasks.

However, while broadly successful, there was evidence to suggest that the benefits of online deployment were not without costs. Data from six participants in the online condition had to be discarded because the participants had made so many errors that their data were not analyzable. That said, with appropriate selection criteria in place, the remaining data were of high quality, which leads us to believe that the issue is one of individual differences rather than a systemic issue with online delivery.

Analysis of performance data suggested that at least one of the six discarded participants was completely inattentive to the task that they were supposed to be performing. The remaining five

participants made too many errors for their data to be usable. We took a close look at the kinds of errors that these participants made. There is some evidence that misunderstandings over the operation of the task may have contributed to the error rates. For instance, participants often tried to return to the last subtask element they completed, rather than the one they were about to work on. Participants were told explicitly to resume at the exact moment they left off from. The literature has shown that repetition of these errors after interruptions are likely (e.g., Altmann et al., 2013; Trafton et al., 2011), but at their rate of occurrence in this group of five participants suggests they may not have fully understood the instructions they were given.
In the lab, an experimenter can quickly correct misunderstandings during training trials, but this was not possible for the online experiment. Participants in the online condition were given the same instructions and completed the same number of training trials as the lab participants. That these misunderstandings occurred suggests that specialized training protocols might be required for online experiments. More interactive training trials, including systems for detecting participants not following instructions (Oppenheimer, Meyvis, & Davidenko, 2009), and training trials that are specifically designed to 'weed-out' participants who have not been able to learn the task quickly enough, could be utilized in future studies to minimize these effects.

It is also possible that the participants who produced insufficient analysable data were switching to other tasks during the experiment and this was sufficiently detrimental to their performance that they fell below the selection thresholds. Indeed, statistical comparisons suggest that these participants took 45 seconds longer, on average, for each trial than other online participants, implying that these rejected participants may have been switching to other tasks while they worked through the experiment. In one instance, a participant took almost ten minutes to complete a single trial. That trials typically took participants around two-and-a-half minutes can only be reasonably explained by interleaving behaviour – these participants must have been working on other tasks while they participated.

Rather than a lack of knowledge, it is also possible that the slowed performance of the online participants arose from the cumulative effects of hardware differences; screen size and input device variations could have affected the speed of completion. However, exploring other quantitative measures of performance in the task suggest that this was not the case.
Measures of error rate and audit accuracy suggest that online participants were just as conscientious as those in the lab. The audit accuracy data are particularly compelling in this regard because there were no direct costs associated with employing a guessing strategy. This suggests that increased trial times were likely the result of distractions appearing in participants' environments, rather than an unwillingness to make an effort to follow instructions. Overall we are confident that the data are sufficiently indistinguishable that they can be treated as equivalent.
Some of our concerns about running the experiment online were realized, but for the most part the problems we had expected did not materialize (e.g., participants dropping out, participants randomly clicking their way through the experiment, technical problems).

This research establishes the viability of online interruption research and in doing so opens a number of avenues of investigation both in the area of interruptions research and in other domains that are concerned with routine procedural action. On the prosaic side, online studies will give experimenters the opportunity to deploy their studies to a large number of participants quickly and cheaply; this will allow for the rapid piloting of ideas for new experiments.

Perhaps more exciting is the opportunity afforded by online studies to investigate interruptions in completely novel ways. For instance, the moments at which participants interleave the experimental task with other activities could be compared with the demands of the task at the moment they switch; one of the difficulties of understanding how people defer interruptions in lab settings is that all of the tasks participants can work on are experimental fabrications, and of little interest to participants. For interruptions research, the pitfalls of online investigation can instead be seen as opportunities to study interruptions and multitasking in a more naturalistic setting.

1.1. Limitations and Generalizability

1.1.1. Limitations

The results of our study highlight at least two difficulties in running time-sensitive experiments in distributed contexts. These issues are sample sizes and participants' discretionary multitasking behaviour outside the experimental task. These limitations may have had some influence on our results and so we discuss each in turn below.

First, relatively small samples were used in this study. This was done to ensure parity of sample sizes in online and lab-based groups. Crowdsourcing using platforms like AMT offers the opportunity to collect large samples. Amongst other things, this gives researchers a chance to examine small effects. Larger samples for both groups would have given an opportunity to have looked even more closely at small variations in behaviour between the two groups.
Second, we were unable to directly measure task switching outside of our task. Activity monitoring studies (e.g., Gould, Cox, & Brumby, 2013; Mao, Kamar, & Horvitz, 2013; Rzeszotarski, Chi, Paritosh, & Dai, 2013) show that in online settings people often do not devote all of their attention to tasks. Being able to monitor these switches is likely to be informative when trying to understand the root cause of variation between results obtained online and in the lab.

1.1.2. Generalizability

We next consider the extent to which the results reported in this paper can be generalized to other settings and might be of relevance to the broader crowdsourcing research community. In our experiment we sampled both groups from the same underlying population; participants were recruited from the same university participant pool. This provides an additional degree of control to the experiment and gives us more confidence in the results. However, although participant

pools have been used in the past (e.g., Dandurand et al., 2008), crowdsourcing platforms like AMT are used more often for online studies of human performance. Do the results of this experiment generalize to populations of paid crowdworkers or volunteer citizen scientists? We contend that our results generalize to other populations. It is well known that multitasking behaviour is influenced by individual differences in many environments (Meys & Sanderson, 2013; Ophir, Nass, & Wagner, 2009; Poposki & Oswald, 2010). The question is whether there are large differences between different populations. Prior work comparing subject pools with AMT workers suggests that for simple studies of human performance, the variation within populations is greater than the variation between populations (see, e.g., Kittur, Chi, & Suh, 2008; Komarov et al., 2013). Paid crowdsourcing platforms do have some peculiarities (e.g., workers are often better educated than the population as a whole), but they are for the most part unremarkable in their makeup (Paolacci & Chandler, 2014).

If our results generalize to other populations, do they generalize to other studies of interruptions? Here the case is not so clear, mostly due to a paucity of evidence. However, the domain of interruptions and multitasking is large and multifaceted. There is a clear dichotomy between routine tasks like the one presented in this study and problem solving tasks that have also been investigated (Hodgetts & Jones, 2006; Hodgetts, Tremblay, Vallières, & Vachon, 2015; P. L. Morgan & Patrick, 2013). Problem solving tasks require participants to be attentive and have a strong conceptual understanding of the problem they are trying to solve. The results of our study suggest that in online settings participants cannot necessarily be relied upon to exhibit these characteristics.

Our results also demonstrate the need to be mindful of the tasks that online participants are given and the support they are given when the task is described. The objectives of this study necessitated changes to the task that led to task behaviour that some participants may have found to be counter-intuitive. Tasks should be redesigned to avoid this situation if possible. Furthermore, training needs to be calibrated for online environments. In the absence of an experimenter, care needs to be taken to ensure that participants have instructions that are easy to assimilate and hard to skip. The use of interactive training trials that monitor behaviour and feedback in real-time would have been a clear improvement for the experiment presented here. These principles hold true for all experiments that require participants to complete a task.

1.2. Conclusion

In this paper we have described a comparative study of interruptions in online and lab settings. Our data showed that data from the two sources could be treated as equivalent, augmenting the findings of previous work with data from a new domain. It also demonstrates that online interruption studies, which are often time-sensitive and subject to strategic decision-making, are viable online. This has implications for researchers in the area as it offers the possibility of novel, naturalistic approaches to interruption research involving routine tasks.

5. REFERENCES

Altmann, E. M., Trafton, G. J., & Hambrick, D. Z. (2013). Momentary Interruptions Can Derail the Train of Thought. *Journal of Experimental Psychology: General*, *143*(1), 215–226. http://doi.org/10.1037/a0030986

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: an activation-based model. *Cognitive Science*, *26*(1), 39–83. http://doi.org/10.1016/S0364-0213(01)00058-1

Araujo, R. M. (2013). 99designs: An Analysis of Creative Competition in Crowdsourced Design. In *First AAAI Conference on Human Computation and Crowdsourcing* (pp. 17–24). AAAI. Retrieved from http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7519

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*(3), 800–813. http://doi.org/10.3758/s13428-011-0081-0

Bernstein, M. S., Brandt, J., Miller, R. C., & Karger, D. R. (2011). Crowds in Two Seconds: Enabling Realtime Crowd-powered Interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (pp. 33–42). New York, NY, USA: ACM. http://doi.org/10.1145/2047196.2047201

Bernstein, M. S., Karger, D. R., Miller, R. C., & Brandt, J. (2012). Analytic Methods for Optimizing Realtime Crowdsourcing. In *Proceedings of Collective Intelligence 2012*. Retrieved from http://arxiv.org/abs/1204.2995

Brumby, D. P., Cox, A. L., Back, J., & Gould, S. J. J. (2013). Recovering from an interruption: Investigating speed-accuracy trade-offs in task resumption behavior. *Journal of Experimental Psychology: Applied*, *19*(2), 95–107. http://doi.org/10.1037/a0032696

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, *6*(1), 3–5. http://doi.org/10.1177/1745691610393980

Cades, D. M., Davis, D. A. B., Trafton, J. G., & Monk, C. A. (2007). Does the Difficulty of an Interruption Affect Our Ability to Resume? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *51*(4), 234–238. http://doi.org/10.1177/154193120705100419

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130. http://doi.org/10.3758/s13428-013-0365-7

Chung, P. H., & Byrne, M. D. (2008). Cue effectiveness in mitigating postcompletion errors in a routine procedural task. *International Journal of Human-Computer Studies*, *66*(4), 217–232. http://doi.org/10.1016/j.ijhcs.2007.09.001

Dabbish, L., Mark, G., & González, V. M. (2011). Why do I keep interrupting myself?: environment, habit and self-interruption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3127–3130). New York, NY, USA: ACM. http://doi.org/10.1145/1978942.1979405

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, *40*(2), 428–434. http://doi.org/10.3758/BRM.40.2.428

Eriksson, K., & Simpson, B. (2010). Emotional reactions to losing explain gender differences in entering a risky lottery. *Judgment and Decision Making*, *5*(3). Retrieved from http://search.proquest.com/docview/1011287632/abstract?accountid=14511

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*(5), 847–857. http://doi.org/10.3758/s13423-012-0296-9

Gilbert, S. J. (2015). Strategic offloading of delayed intentions into the external environment. *The Quarterly Journal of Experimental Psychology*, *68*(5), 971–992. http://doi.org/10.1080/17470218.2014.972963

González, V. M., & Mark, G. (2004). 'Constant, constant, multi-tasking craziness': managing multiple working spheres. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 113–120). New York, NY, USA: ACM. http://doi.org/10.1145/985692.985707

Gould, S. J. J., Brumby, D. P., & Cox, A. L. (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), 149–153. http://doi.org/10.1177/1541931213571034

Gould, S. J. J., Cox, A. L., & Brumby, D. P. (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* (pp. 22–23). Retrieved from http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7485

Gray, W. D. (2000). The nature and processing of errors in interactive behavior. *Cognitive Science*, *24*(2), 205–248. http://doi.org/10.1016/S0364-0213(00)00022-7

Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 203–212). New York, NY, USA: ACM. http://doi.org/10.1145/1753326.1753357

Hodgetts, H. M., & Jones, D. M. (2006). Contextual Cues Aid Recovery From Interruption: The Role of Associative Activation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1120–1132. http://doi.org/10.1037/0278-7393.32.5.1120

Hodgetts, H. M., Tremblay, S., Vallières, B. R., & Vachon, F. (2015). Decision support and vulnerability to interruption in a dynamic multitasking environment. *International Journal of Human-Computer Studies*, *79*, 106–117. http://doi.org/10.1016/j.ijhcs.2015.01.009

Hsieh, G., Kraut, R. E., & Hudson, S. E. (2010). Why Pay?: Exploring How Financial Incentives Are Used for Question & Answer. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 305–314). New York, NY, USA: ACM. http://doi.org/10.1145/1753326.1753373

Jennett, C., Furniss, D. J., Iacovides, I., Wiseman, S., Gould, S. J. J., & Cox, A. L. (2014). Exploring Citizen Psych-Science and the Motivations of Errordiary Volunteers. *Human Computation*, *1*(2). http://doi.org/10.15346/hc.v1i2.10

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 453–456). New York, NY, USA: ACM. http://doi.org/10.1145/1357054.1357127

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., … Horton, J. (2013). The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 1301–1318). New York, NY, USA: ACM. http://doi.org/10.1145/2441776.2441923

Komarov, S., Reinecke, K., & Gajos, K. Z. (2013). Crowdsourcing Performance Evaluations of User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 207–216). New York, NY, USA: ACM. http://doi.org/10.1145/2470654.2470684

Lasecki, W. S., & Bigham, J. P. (2012). Online quality control for real-time crowd captioning. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility* (pp. 143–150). New York, NY, USA: ACM. http://doi.org/10.1145/2384916.2384942

Lasecki, W. S., Rzeszotarski, J. M., Marcus, A., & Bigham, J. P. (2015). The Effects of Sequence and Delay on Crowd Work. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1375–1378). New York, NY, USA: ACM. http://doi.org/10.1145/2702123.2702594

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., … Vandenberg, J. (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, *389*(3), 1179–1189. http://doi.org/10.1111/j.1365-2966.2008.13689.x

Li, S. Y. W., Blandford, A., Cairns, P., & Young, R. M. (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), 314–328. http://doi.org/10.1037/a0014397

Li, S. Y. W., Cox, A. L., Blandford, A., Cairns, P., & Abeles, A. (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* (pp. 471–476). Vancouver, BC, Canada: Cognitive Science Society.

Mao, A., Kamar, E., & Horvitz, E. (2013). Why Stop Now? Predicting Worker Engagement in Online Crowdsourcing. In *First AAAI Conference on Human Computation and Crowdsourcing* (pp. 103–111). AAAI. Retrieved from http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7498

Mark, G., Gonzalez, V. M., & Harris, J. (2005). No task left behind?: examining the nature of fragmented work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 321–330). New York, NY, USA: ACM. http://doi.org/10.1145/1054972.1055017

Mark, G., Gudith, D., & Klocke, U. (2008). The cost of interrupted work: more speed and stress. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 107–110). New York, NY, USA: ACM. http://doi.org/10.1145/1357054.1357072

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23. http://doi.org/10.3758/s13428-011-0124-6

Mason, W., & Watts, D. J. (2010). Financial incentives and the 'performance of crowds'. *SIGKDD Explor. Newsl.*, *11*(2), 100–108. http://doi.org/10.1145/1809400.1809422

Meys, H. L., & Sanderson, P. M. (2013). The Effect of Individual Differences on How People Handle Interruptions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 868–872. http://doi.org/10.1177/1541931213571188

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

Moore, J., Gay, P. L., Hogan, K., Lintott, C., Impey, C., & Watson, C. (2011). Facebooking Citizen Science with the Zooniverse. In *Bulletin of the American Astronomical Society* (Vol. 43, p. 15813). Retrieved from http://adsabs.harvard.edu/abs/2011AAS...21715813M

Morgan, B., & D'Mello, S. (2013). The Effect of Positive vs. Negative Emotion on Multitasking. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 848–852. http://doi.org/10.1177/1541931213571184

Morgan, P. L., & Patrick, J. (2013). Paying the price works: Increasing goal-state access cost improves problem solving and mitigates the effect of interruption. *The Quarterly Journal of Experimental Psychology*, *66*(1), 160–178. http://doi.org/10.1080/17470218.2012.702117

Newell, A., & Card, S. K. (1985). The prospects for psychological science in human-computer interaction. *Human-Computer Interaction*, *1*(3), 209–242. http://doi.org/10.1207/s15327051hci0103_1

Ophir, E., Nass, C., & Wagner, A. D. (2009). Cognitive control in media multitaskers. *Proceedings of the National Academy of Sciences*, *106*(37), 15583–15587. http://doi.org/10.1073/pnas.0903620106

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. http://doi.org/10.1016/j.jesp.2009.03.009

Paolacci, G., & Chandler, J. (2014). Inside the Turk Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, *23*(3), 184–188. http://doi.org/10.1177/0963721414531598

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.

Payne, S. J., Duggan, G. B., & Neth, H. (2007). Discretionary task interleaving: Heuristics for time allocation in cognitive foraging. *Journal of Experimental Psychology: General*, *136*(3), 370–388. http://doi.org/10.1037/0096-3445.136.3.370

Poposki, E. M., & Oswald, F. L. (2010). The Multitasking Preference Inventory: Toward an Improved Measure of Individual Differences in Polychronicity. *Human Performance*, *23*(3), 247–264. http://doi.org/10.1080/08959285.2010.487843

Ratwani, R. M., & Trafton, J. G. (2008). Spatial memory guides task resumption. *Visual Cognition*, *16*(8), 1001–1010. http://doi.org/10.1080/13506280802025791

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (pp. 2863–2872). New York, NY, USA: ACM. http://doi.org/10.1145/1753846.1753873

Rzeszotarski, J. M., Chi, E., Paritosh, P., & Dai, P. (2013). Inserting Micro-Breaks into Crowdsourcing Workflows. In *Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts AAAI Technical Report CR-13-01* (pp. 62–63). AAAI. Retrieved from http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7544

Rzeszotarski, J. M., & Kittur, A. (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* (pp. 13–22). New York, NY, USA: ACM. http://doi.org/10.1145/2047196.2047199

Rzeszotarski, J. M., & Kittur, A. (2012). CrowdScape: interactively visualizing user behavior and output. In *Proceedings of the 25th annual ACM symposium on User interface software and technology* (pp. 55–62). New York, NY, USA: ACM. http://doi.org/10.1145/2380116.2380125

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

Salvucci, D. D., & Bogunovich, P. (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* (pp. 85–88). New York, NY, USA: ACM. http://doi.org/10.1145/1753326.1753340

Shaw, A. D., Horton, J. J., & Chen, D. L. (2011). Designing Incentives for Inexpert Human Raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (pp. 275–284). New York, NY, USA: ACM. http://doi.org/10.1145/1958824.1958865

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (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* (pp. 254–263). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1613715.1613751

Stanton, J. M. (1998). An Empirical Assessment of Data Collection Using the Internet. *Personnel Psychology*, *51*(3), 709–725. http://doi.org/10.1111/j.1744-6570.1998.tb00259.x

Suri, S., & Watts, D. J. (2011). Cooperation and Contagion in Web-Based, Networked Public Goods Experiments. *PLoS ONE*, *6*(3). http://doi.org/10.1371/journal.pone.0016836

Trafton, J. G., Altmann, E. M., & Ratwani, R. M. (2011). A memory for goals model of sequence errors. *Cognitive Systems Research*, *12*(2), 134–143. http://doi.org/16/j.cogsys.2010.07.010