Assessing the Viability of Online Interruption Studies

Sandy J. J. Gould, Anna L. Cox, Duncan P. Brumby, Sarah Wiseman

UCL Interaction Centre, University College London, Gower Street, London, UK
s.gould@cs.ucl.ac.uk, anna.cox@ucl.ac.uk, brumby@cs.ucl.ac.uk, s.wiseman@cs.ucl.ac.uk

Abstract

Researchers have been collecting data online since the early days of the Internet and as technology improves, increasing numbers of traditional experiments are being run online. However, there are still questions about the kinds of experiments that work online, particularly over experiments with time-sensitive performance measures. We are interested in one time-sensitive measure specifically, the time taken to resume a task following an interruption. We ran participants through an archetypal interruption study online and in the lab. Statistical comparisons showed no significant differences in the time it took to resume following an interruption. However, there were issues with data quality that stemmed from participant confusion about the task. Our findings have implications for experiments that assess time-sensitive performance measures in tasks that require continuous attention.

Introduction

Interruptions are disruptive, reducing performance and increasing error rates. Understanding the effects of interruption and developing techniques for mitigating their effects has been an active area of research in the fields of human-computer interaction and human factors. Experimental investigations of interruptions are often time consuming to conduct and might benefit from the scale and cost-effectiveness of online experimentation.

In many other areas of research, online data collection is an increasingly popular way of running experiments quickly and cheaply (e.g., Kittur, Chi, and Suh 2008; Suri and Watts 2011). However, experimental paradigms are not homogeneous; their appearance, design and measures vary considerably. These factors might make a particular paradigm more or less suitable for online deployment. Although there are obvious concerns about the lack of control that can manifest when experiments are run online, comparative studies of online and lab-collected data have often found no statistically significant differences in performance (e.g., Komarov, Reinecke, and Gajos 2013; Dandurand, Shultz, and Onishi 2008). Nonetheless, experiments vary in the degree of control that they require and studies to date have tended to use short, simple tasks in which potential confounds, such as multitasking behavior, are naturally minimized.

In this paper we focus on studies of interruption, which typically require participants to work on long, complex tasks that require continuous attention and actively encourage or mandate disruptive multitasking. Given these characteristics, we wondered whether online studies of interruption were viable. To address this question, we ran an experiment online and in the lab. We found that there was no significant difference between online and lab data. This work contributes, to our knowledge, the first online study of interruption and a comparison of lab-collected and online data in a relatively long-lasting, time-sensitive experiment that requires sustained attention.

Method

A total of 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 a university subject pool and were paid £7 for approximately one hour of their time.

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. There was one dependent variable, *resumption lag* (Altmann and Trafton 2002), which is the time it took participants to resume working on the primary task after being interrupted.

Our primary focus in this study was the extent to which the delivery medium of the study affected participants' performance. Although the measures we use are appropriate to the field of interruption research, a discussion of the results in the context of existing interruption literature is outside the scope of this paper.

The task in this experiment was the *Pharmacy Task*, an adaptation of the *Doughnut Machine* (Li et al. 2006). This is a routine data-entry task that has been used previously to investigate the effects of interruptions. Participants are given a set of 'prescriptions' that contain values that must be copied into one of the five subtasks that make up the task. From time-to-time, participants were interrupted. When participants resuming after an interruption, any cues in the task that might aid resumption were removed. After completing the experiment, participants were given a twelve-item questionnaire, which asked about their experience of the task and the interruptions.

Results

The primary measure was resumption lag, or the time it took participants to resume after an interruption. There was no significant main effect of experiment location (online vs. lab), $F(1,39)=0.02$, $p=0.88$, or of interruption timing, $F(1,39)=4.00$, $p=.053$. There was however a significant main effect of interruption relevance, $F(1,39)=9.62$, $p<.01$, $\eta_p^2=.20$. There were four interactions (location × type × timing; location × type; type × timing; location × timing); none of which were significant (i.e., $p > .05$). In addition to the resumption lag data, we also recorded whether participants resumed the task in the correct place. Error rates were approximately the same regardless of participant location, with online participants resuming incorrectly 14% of the time compared with 16% for online participants.

Discussion

Our results showed no statistically significant differences between the data produced by participants who took part online and those who participated in the lab. This demonstrates the viability of online interruption research for the first time in an experiment that is time consuming, sensitive to strategy, and generates relatively small quantities of data. This finding also augments previous comparative work in the area that has come to similar conclusions (e.g., Komarov, Reinecke, and Gajos 2013; Dandurand, Shultz, and Onishi 2008; Paolacci, Chandler, and Ipeirotis 2010).

However, while broadly successful, there was evidence to suggest that the benefits of online deployment were not without cost. Data from six participants in the online condition had to be discarded because the participants had resumed incorrectly in all trials for some conditions, rendering their data unusable. That said, with appropriate selection criteria in place (i.e., resumption lags <±1.96< SDs of mean), the remaining data were of high quality, leading us to believe that the issue was one of individual variation rather than systematic issues with online delivery.

Establishing the viability of online interruption research opens a number of avenues of investigation for future work. On the prosaic side, online studies will give researchers 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 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 work on are fabricated, and consequently of little interest to participants. For interruption research, the pitfalls of online investigation can instead be seen as opportunities to study interruptions and multitasking in a naturalistic setting.

Acknowledgements

Work funded by CHI+MED, EPSRC grant EP/G059063/1.

References

Altmann, Erik M., and J. Gregory Trafton. 2002. 'Memory for Goals: An Activation-based Model'. *Cognitive Sci* 26 (1): 39–83.

Dandurand, Frédéric, Thomas R. Shultz, and Kristine H. Onishi. 2008. 'Comparing Online and Lab Methods in a Problem-solving Experiment'. *Behav Res Methods* 40 (2): 428–434.

Kittur, Aniket, 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*, 453–456.

Komarov, Steven, 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*.

Li, Simon Y. W., 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*, 471–476.

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

Suri, Siddharth, and Duncan J. Watts. 2011. 'Cooperation and Contagion in Web-Based, Networked Public Goods Experiments'. *PLoS ONE* 6 (3) (March 11).