Northumbria Research Link

The Effects of Simulated Interruptions on Mobile Search Tasks

Orland Hoeber[1], Morgan Harvey[2], Shaheed Ahmed Dewan Sagar[1],
and Matthew Pointon[4]

[1]Department of Computer Science, University of Regina, Canada
[2]Information School, The University of Sheffield, UK
[3]Department of Computer and Information Sciences, Northumbria University, UK

Abstract

While it is clear that using a mobile device can interrupt real-world activities such as walking or driving, the effects of interruptions on mobile device use have been under-studied. We are particularly interested in how the ambient distraction of walking while using a mobile device, combined with the occurrence of simulated interruptions of different levels of cognitive complexity, affect web search activities. We have established an experimental design to study how the degree of cognitive complexity of simulated interruptions influence both objective and subjective search task performance. In a controlled laboratory study (n=27), quantitative and qualitative data were collected on mobile search performance, perceptions of the interruptions, and how participants reacted to the interruptions, using a custom mobile eye-tracking app, a questionnaire, and observations. As expected, more cognitively complex interruptions resulted in increased overall task completion times and higher perceived impacts. Interestingly, the effect on the resumption lag or the actual search performance was not significant, showing the resiliency of people to resume their tasks after an interruption. Implications from this study enhance our understanding of how interruptions objectively and subjectively affect search task performance, motivating the need for providing explicit mobile search support to enable recovery from interruptions.

Keywords: mobile search, interruptions, controlled laboratory study

Introduction

In recent years, web search has changed rapidly from something done almost exclusively in front of a computer to something that we often do with mobile devices in a

---

variety of mobile situations. Since around 2015, more searches are performed on mobile devices than on desktop or laptop computers (Dischler, 2015). Real-world mobile tasks occur in many different environments, such as on public transport, while walking from place to place (Lin et al., 2007) or in social contexts (Church & Oliver, 2011), in which there can be many other things vying for the user's attention. These can break the user's thought process, dividing their attention, and often resulting in considerable interruption before they can resume their original task (Nicolau & Jorge, 2012).

A substantial amount of research has been conducted on the impact of mobile device use on other activities, such as driving (Caird et al., 2008; Larsen et al., 2020; Oviedo-Trespalacios et al., 2016), but relatively little has been conducted on the effects of environmental interruptions on mobile device use. What has been done has either been focused on the impact of constant walking (continuous mild distraction) on searchers' objective performance and on their perceptions of the task (Harvey & Pointon, 2017, 2019); or has involved users following a route through a city, where many variables remain uncontrolled and thus internal validity is low (Oulasvirta et al., 2005).

What remains unclear is how explicit interruptions that occur during the stages of a mobile search process (querying, evaluating the search results, reading the documents) affect the searcher and their performance. More specifically, it is unknown how the cognitive complexity of an interruption affects the ability of mobile searchers to resume and complete their search activities, and how it influences their subjective impressions of its impact. To address this gap in the literature, this study was guided by two core research questions:

RQ1: How do interruptions of differing levels of cognitive complexity affect mobile search task performance?

RQ2: How do interruptions of differing levels of cognitive complexity affect the subjective impressions of the degree/impact of the interruptions and engagement in the search task?

To answer these research questions we designed a controlled laboratory study to introduce simulated, explicit, and repeatable interruption events of varying levels of cognitive complexity during specific stages of mobile search activities. We took steps to ensure that the experiences of the participants were as similar to one another as possible (e.g., the study was conducted in a controlled environment; all participants walked on a treadmill at a comfortable pace), allowing us to study causality (Kelly, 2009). We developed a custom iPhone app that tracks a participant's gaze while they look at the device, and an automated logging system that detects when the gaze leaves the device, when it returns, and when interaction with the device resumes. This app can be configured to load any web site; for the purposes of this study we use the Google search interface. We measured the duration of each interruption, the time it took for the participants to return to their search tasks, the total time to task completion, selected document relevance, and the participants' perceptions of the degree and impact of the interruption on the different stages of their search tasks. We also observed the participants' behaviours immediately before and after the interruptions and documented these in field notes.

This research was informed by our prior work that studied the impact of walking on search performance (Harvey & Pointon, 2017, 2019), the conduct of controlled laboratory

studies of mobile search interfaces (Gopi & Hoeber, 2016), and the design of interfaces to support search task resumption (Gomes & Hoeber, 2021). While research in cognitive psychology has sought to understand the impact of interruptions on cognitive processes (Couffe & Michael, 2017; Lavie, 2005), such studies have been conducted using tightly-controlled environments and activities. This study extends such research, using realistic distracting/interrupting elements that are both physical (walking) and cognitive (searching) in nature.

Our research is novel both in focus and approach; we are unaware of any other studies that use mobile search as the baseline and add experimental conditions for different levels of interruptions, nor controlled laboratory studies that place searchers on a treadmill and use the front-facing camera of a mobile device to detect when attention leaves and returns to the device. Results from this study contribute to our knowledge of the objective and subjective effects of interruptions on mobile search task stages. It provides a basis for the future design of novel mobile search interfaces that mitigate the effects of interruptions, allowing searchers to more easily resume their tasks after an interruption.

Literature Review

An important precursor to establishing the literature that motivates this research is to explain the differences between distractions and interruptions. While much of the literature uses these terms interchangeably, there is a distinct difference that is relevant to our work. Both distractions and interruptions start with a lapse of attention from the primary task, caused by some event. A distraction itself does not always result in active engagement in the event. When engagement does occur, it causes an interruption of the primary task (Couffe & Michael, 2017). So, while walking on the street and performing a mobile search, one might be distracted by other people and traffic, but not be interrupted. However, if one of these distractions becomes significant enough to draw the searcher's attention away from their device and on to the distraction (e.g., someone asking for directions), then this becomes an interruption.

To set the context for our work and motivate the specific research questions and methodological decisions taken, we summarize three bodies of related work: (1) research on Information Retrieval (IR) in a mobile context and how this differs from searches performed in a more traditional desktop context; (2) recent research on the impact of distractions and interruptions on mobile device use and mobile search; and (3) foundations from cognitive psychology for simulating interruptions within a laboratory setting.

Mobile Information Retrieval

Recent years have seen a dramatic increase in the number of search queries originating from mobile devices and, consequently, an increase in the amount of literature that investigates web search in a mobile context (Crestani et al., 2017). Early work analysed query logs to understand the characteristics of mobile search queries and how these differ from their desktop browser-based counterparts (Church et al., 2008; Kamvar & Baluja, 2006). Kamvar and Baluja (2006) analysed over 1 million hits to Google's early mobile search site and found that mobile queries were similar in length to desktop queries but noted that the amount of effort required to input text was demonstrably greater. Later

work by Church et al. (2008) found that mobile queries were actually slightly shorter, that the topical distribution for mobile searches was different, and that many more queries were abandoned with zero result clicks.

Much work has identified the increased importance of situational context for mobile search (Aliannejadi et al., 2019; Sohn et al., 2008; Teevan et al., 2011). A diary study by Sohn et al. (2008) found that the majority of mobile queries (72%) were prompted by contextual factors and that, in more than a third of cases, the most crucial contextual element was location, although social context, time, and activity were also important factors. Using a similar methodology, Teevan et al. (2011) investigated context in lower granularity and found that, in 68% of cases, their participants were searching while in transit and that 63% of searches were in a social context. More recent work (Aliannejadi et al., 2019) captured data from a task-based field study with 31 participants, providing more objective data on the importance of context for mobile search. The results suggest that context has a large impact on how people go about performing tasks and that searching on-the-go, be it on foot or on transport, is not only common but has significant impacts on search behaviour and performance.

Distractions, Interruptions, and Mobile Search

Much of the research on distractions and interruptions of search activities highlights an obvious difference between desktop and mobile search. Desktop search may be subject to interruptions that lead to multi-session search behaviour for which novel search interfaces may be designed to support task resumption (Gomes & Hoeber, 2021; Morris et al., 2008). By contrast, mobile search often takes place in situations where there is great potential for users to be distracted from their search task by elements in the surrounding environment. When these distractions become interruptions, they typically lead to reduced performance when interacting with the device's user interface (Lin et al., 2007), resulting in slower text input speeds (Mizobuchi et al., 2005), shorter queries, and an increased probability of typographical errors (Schaller et al., 2012). Walking and interacting at the same time requires a user to divide their attention between the two tasks (Lamberg & Muratori, 2012) and has effects that extend beyond just interaction with the interface. Reading comprehension can be significantly reduced (Barnard et al., 2007) and people's perception of their own performance and of the search task can be negatively impacted (Harvey & Pointon, 2017, 2019).

Prior work has investigated the effects of common distractions and their resulting interruptions both in-the-wild (Hoggan et al., 2009; Oulasvirta et al., 2005) and via laboratory-based experiments (Brumby et al., 2013; Harvey & Pointon, 2017, 2019; Nicolau & Jorge, 2012). Oulasvirta et al. (2005) had participants complete tasks while following a pre-defined, but otherwise uncontrolled, route through a city and compared their behaviour with those completing tasks in a lab. The city-navigating participants experienced significant impairment when compared to those in the lab caused by their need to much more rapidly and frequently divide their attention between their assigned task and the environment. Later work (Hoggan et al., 2009) demonstrated that noisy and bumpy environments have a similar effect, reducing the ability of users to accurately interact with the device and increasing their interaction times.

Controlled laboratory-based experiments (Harvey & Pointon, 2017) have shown that fragmented attention of users while searching on-the-go affects objective search performance along with users' perception of task difficultly and of their own performance. More detailed analysis by the same authors (Harvey & Pointon, 2019) revealed that ambulatory motion had a negative impact on the ability of participants to read documents and assess them for relevance and increased the amount of time necessary to correct query input errors. Brumby et al. (2013) simulated cognitively demanding secondary mental arithmetic tasks demonstrating that these resulted in users erring when they subsequently resumed their primary data entry tasks. More recent work (Sarsenbayeva et al., 2018) has shown that being exposed to speech and urban noise significantly slows down text entry, although it is unclear whether this would have a similarly dramatic effect on web search tasks.

Larsen et al. (2020) investigated whether the use of conversational agents (e.g. Apple's Siri) reduces the impact of distractions when driving compared to manual interaction. Although the work does not seek to specifically investigate the impact of distractions on driving, the authors found that drivers were sufficiently distracted by this interaction paradigm to the extent that such interaction would still be unsafe. Other recent work (Bâce et al., 2020) sought to quantify the attention of users towards their mobile devices and found that using a device in more distracting environments (e.g. when walking or using public transport) results in considerably lower duration of sustained visual attention, implying that users are more often being interrupted from performing tasks on their mobile devices.

Simulating Interruptions

When presented with a distraction, whether we attend to it or not depends on our current perceptual and cognitive load (Lavie, 2005). Since our perceptual processing of the world around us happens automatically, we cannot choose to not perceive what we see and hear, short of closing our eyes and plugging our ears. However, if our perceptual capacity is fully engaged, we may not observe the distraction; likewise, if our cognitive capacity is fully engaged, even if the distraction is perceived, we may not notice it. If we do notice the distraction and wish to ignore it, it is necessary to exert cognitive effort to do so, resulting in what is called selective attention.

When a distraction becomes an interruption (due to active engagement in the event), the degree of impact on the primary task is based on specific factors of the interruption and the effort in transitioning back to the primary task. Wobbrock (2019) suggests that situations, contexts, or environments can negatively affect the ability of people to interact with technology and, as such, can be approached from an accessibility point-of-view. These are referred to as situationally-induced impairments and disabilities (SIIDs) and can induce similar error rates by able-bodied individuals using mobile devices in distracting situations as physically-impaired people using a desktop computer (Yesilada et al., 2010). Such SIIDs can be caused by numerous situation factors, including being interrupted by nearby conversations (Mayer et al., 2018) or other loud noises (Sarsenbayeva et al., 2018; Wolf et al., 2017).

Such impairments induced by interruptions are a result of the re-allocation of attention and interference with one's working memory (Couffe & Michael, 2017). If the action required by the interruption is routine, a minimal amount of attention and working memory will be needed to attend to the interruption; if, however, the action requires recall and

deliberate processing, a non-trivial amount of attention and working memory will need to be re-allocated and there will be interference with working memory. Furthermore, if the interruption is very similar to the target task, the interference with the working memory will be greater than if the interruption is different in nature (Wickens, 2008). These affect the resumption lag: the time between ending the interruption task and returning to the primary task, during which one redirects their attention back to the task, reacquaints themselves with what they were doing, and decides what to do next.

A final consideration is when the interruption occurs in relation to the task steps and substeps. Any reasonably complex task can be decomposed into global steps which are made up of many smaller substeps. For web search tasks, we consider these global steps to be the stages of query formulation/execution, examination of the search results, and examination of an individual document (adapted from Gwizdka (2010)). Research has found that interruptions between the global steps represent course breakpoints in the cognitive activity, and are less damaging than interruptions that occur between or during the substeps (Adamczyk & Bailey, 2004).

What this tells us for simulating interruptions in a laboratory environment is that: (1) participants must not have their full perceptual capacity engaged; (2) the interruption event must be sufficiently different from the target activity to be perceived; (3) participants must not have their full cognitive capacity engaged; and (4) participants must not attempt to actively ignore the interruption event using selective attention. For the interruptions themselves, we must be mindful of the differences in the cognitive affect of the desired action. That is, we must consider: (1) the attention and working memory load of the action; (2) the degree of working memory interference due to similarities between the interruption action and the primary task; and (3) the need to re-allocate attention to deal with the action. Furthermore, we must consider when the interruption occurs during the steps and substeps of the primary task. Controlling for these complexities through careful laboratory study design and the simulation of interruptions ensures that we do not introduce confounding factors.

Methodology

To study the effects of interruptions on mobile search tasks, a controlled laboratory study was conducted in a style that Kelly (2009) classified as an "experimental information behavour" study. We recognize that any empirical study must make trade-offs between the benefits of control against ecological validity. As much as possible, variables not under direct investigation were normalized (e.g., comfortable walking pace, two-handed device use, task complexity, no unplanned interruptions), allowing us to control the study situation and isolate the impact of the variables of interest (e.g., the cognitive complexity of the interruption) (Kelly, 2009). The activity of walking while using a mobile device is itself an ambient distraction from the primary task of using the device. However, there is evidence that experienced mobile device users are good at choosing a walking speed that does not significantly impact their ability to use the device (Bergstrom-Lehtovirta et al., 2011). Variables related to individual differences among the participants (e.g., physical abilities such as typing speed; cognitive abilities such as short-term memory and reading speed) were not controlled in this study. While we expect that such differences may impact search performance, such differences are minor compared to the experimental conditions in this

study, and therefore may be treated as noise in the data. While such studies represent an abstraction and simplification of the real world and, therefore, may not be broadly generalized, the removal of potentially confounding factors permits direct comparisons to be made between participants, and causal relationships to be inferred between the manipulation of the independent variables and the data collected on the dependant variables.

Simulated Interruptions

Using a within-subjects design, one independent variable was manipulated: the cognitive complexity of the interruptions. As the literature suggests that the degree of the interruption depends on how much attention and working memory need to be re-allocated to the interruption (Couffe & Michael, 2017), we used this as the guiding factor in the design of three different levels of simulated interruption: no interruption, low interruption, and high interruption. At the *no interruption* level, we allowed the participants to conduct the search task without any explicit interruptions, providing a baseline of walking and searching. At the *low interruption* level, we interrupted the participants with simple shape recognition activities, requiring a small re-allocation of attention and working memory to perform this simple recall task. At the *high interruption* level, we interrupted the participants with a task of performing a mathematical calculation involving a variable related to their own life (e.g., where x is the last digit of your phone number), which required a moderate degree of attention and working memory re-allocation. This mirrors the approach taken to reliably induce moderate interruptions in previous research (e.g., Brumby et al., 2013).

Of note is that all interruptions were based on external events that we controlled, rather than internally-induced interruptions (such as remembering to do something). We did not exceed the moderate level in terms of attention and working memory re-allocation, as going to such an extreme runs the risk of producing a full task switch, which would no longer be considered merely an interruption. While it would be possible to create more than three levels of this independent variable, we chose to keep this number small to avoid extending the study time and having fatigue become a confounding factor.

The simulated interruptions themselves were intentionally created to deviate from real-world interruptions, providing abstract activities that allowed us to control the degree to which the interruptions required re-allocation of attention and working memory. Even so, they were designed to mimic the cognitive complexity of interruptions in real-world activities of walking and searching (no interruption), a routine interruption such as looking for a walk light at an intersection (low interruption; minor attention and working memory re-allocation), and a more cognitively complex interruption such as being asked how long it will take to get to a specific destination (high interruption; moderate attention and working memory re-allocation). These place cognitive load on the participants in the same way as real-world interruptions would; from this perspective they can be considered abstract representations of what happens in reality. This approach is common in psychology-based experimental design, trading realism for the ability to control bias and avoid the introduction of confounding factors, resulting in an increase to internal validity at the expense of ecological validity. Of note, we explicitly avoided creating any interruption activities that were similar to the target tasks (typing, reading, and assessing relevance) to avoid the known effect that too much similarity between the main task and the interruption can have a very large impact due to working memory interference (Wickens, 2008).

Level	Cognitive impact	Real world examples	Interruption instances
none/no	no re-allocation of attention or working memory (baseline)	walking down an empty street	
low	minor re-allocation of attention and working memory	looking up at a somewhat complex traffic light at an intersection; someones shouts loudly across the street	*What is the name of this shape?* (square) / *What is the name of this shape?* (circle) / *What is the name of this shape?* (triangle)
high	moderate re-allocation of attention and working memory	someone stops you to ask for directions to downtown or to sign a petition	If **x** is the **number of people in your family**, what is the result of this calculation? $\frac{(x+3)3}{2}$ / If **x** is the **number of letters in your first name**, what is the result of this calculation? $\frac{(x-2)3}{2}$ / If **x** is the **last digit in your phone number**, what is the result of this calculation? $\frac{(x+2)3}{2}$

Table 1

Summary of the interruptions used in this study.

For each of the low and high interruption levels, three similarly complex interruption activities were created, as shown in Table 1. When these interruptions occurred, a short explosion noise was played (duration of two seconds), the interruption was shown on an external display screen positioned directly in front of the participant, and a voice-over reading the interruption activity details was played. The purpose of the explosion noise was to alert the participant and ensure that the event was actually perceived (Lavie, 2005), causing attention to be re-allocated and the interruption to occur. Given that the explosion noise and interruption activity happen concurrently, we consider the noise to be intrinsic to the interruption being induced rather than serving as its own independent interruption. The purpose of the voice-over was to mitigate differences in reading speed among the participants, and to make the activity sufficiently different from the target task of searching. Participants were instructed to look up at the screen immediately upon hearing the explosion sound, and to verbally respond to the question presented before continuing with their search task. For the no interruption level, a blank screen was shown on the external display throughout the search process, and no noises were played.

During a particular search task, the interruptions were initiated at three different times: two seconds after the participant had started to type in the query, two seconds after the participant had begun evaluating the search results list, and four seconds after the participant started reading the first search result document. This timing was determined based on observations of our own mobile search behaviour, and ensured that the interruptions occurred in the midst of the search task stages of querying, examining the search results, and examining an individual document (rather than between them, where the literature suggests they will have less impact (Adamczyk & Bailey, 2004; Couffe & Michael, 2017)). Note that this represents a subset of the normal web search information seeking process (Gwizdka, 2010), with a focus on what happens beginning with the query entry stage. That is, this study does not address query formulation, since it would be difficult to observe in a controlled laboratory setting and with our data collection instruments. In order to avoid adding complexity to the study design, the interruptions used at the three stages of the search process were all from the same level, in the order shown in Table 1.

While the process of interrupting the participants multiple times during the search tasks may have allowed some participants to learn the structure of the study, we explicitly asked them to not try to anticipate when an interruption might occur, but instead to focus on the assigned tasks. As expecting an interruption is not uncommon in real-world mobile search activities (e.g., while walking in a busy hallway at a University, where one might expect to run into their friends), we do not view this aspect of the study design as an unreasonable limitation.

Search Tasks

Potential task-based learning effects were managed by having participants perform a different task for each of the three interruption levels (treating the task as a between-subjects variable). These tasks were chosen from the most recent TREC 2014 Web Track, which provides a set of carefully-developed test topics [1] designed to go beyond simple fact verification, requiring the searcher to consider the relevance of the search results. The

_____

[1]https://trec.nist.gov/data/web/2014/trec2014-topics.xml

three tasks were chosen to have similar length two-word queries, and focus on topics that can be generally understood but are not part of common knowledge (see Table 2). The chosen tasks are informational in nature, which are consistent with the majority of mobile search tasks (Church & Smyth, 2009), and involve fact-finding, which is also common in this context (Carrascal & Church, 2015). They are, therefore, broadly representative of tasks one might typically perform in a mobile setting. In particular, we can consider these tasks as the type that may be conducted to address an information need that could occur within a discussion among friends, or a fleeting situational information need while the searcher is on the go. They were intentionally chosen to be rather obscure, to minimize the chance that a participant might have conducted a similar search recently or would already be familiar with suitable answers. The chosen tasks are also not either location- or culture-dependent, which could have otherwise introduced additional biases, particularly as our participants were from different countries. Participants were instructed to use the provided query, which was displayed to them clearly on the screen and underlined, and to not use the query completion feature provided by the search engine.

The order of exposure of the interruption levels (3) and tasks (3) was rotated and counter-balanced using a Graeco-Latin square, producing nine different permutations of task and interruption orders. This method ensured that any order and learning effects (including learning to anticipate interruptions) were distributed evenly across the data.

Study Setup

The physical setup of the study consisted of: a treadmill, on which participants were instructed to walk while performing the assigned search tasks; a platform, on which the researchers could stand to observe the participants from behind; a computer and external display screen in front of the participants, on which to show both general instructions for each of the tasks and the interruptions. A remote was used to allow the researchers to initiate the interruptions during each of the stages of the search task; and a mobile device (iPhone X) was given to the participant to use for the search tasks (see Figure 1). Each participant was given specific instructions on using the device with two hands in order to

Task	ID	Task Description	Query	Fam.	Int.	Diff.	Dur.
training	294	What are the names of some flowering plants?	"flowering plants"				
task 1	297	How can one prevent altitude sickness?	"altitude sickness"	2.63	2.59	1.78	78 s
task 2	272	Find data on how to generally interpret dreams.	"interpret dreams"	1.74	2.56	2.41	59 s
task 3	290	How do you identify a Norway Spruce?	"norway spruce"	1.51	2.07	2.11	55 s

Table 2

The tasks assigned to the participants in this study, their pre-task mean responses to questions regarding familiarity, interest, and expected difficulty (5=extremely; 3=moderately; 1=not at all), and task duration (no interruption condition).

minimize the amount of movement of the device in relation to the participant's torso and head. If the participant was naturally a two-handed user, they could use the device as normal; if they were normally a one-handed user, the instruction was to simply add the second hand to the device to stabilize it. While such two-handed use may have not matched the normal mode of use for some participants, it is the most common way in which mobile devices are used while walking (Harvey & Pointon, 2019), and was found to increase the accuracy of the eye tracking during pilot testing. Doing so also served to remove a potential confounding factor from the study design: one-handed vs. two-handed use.
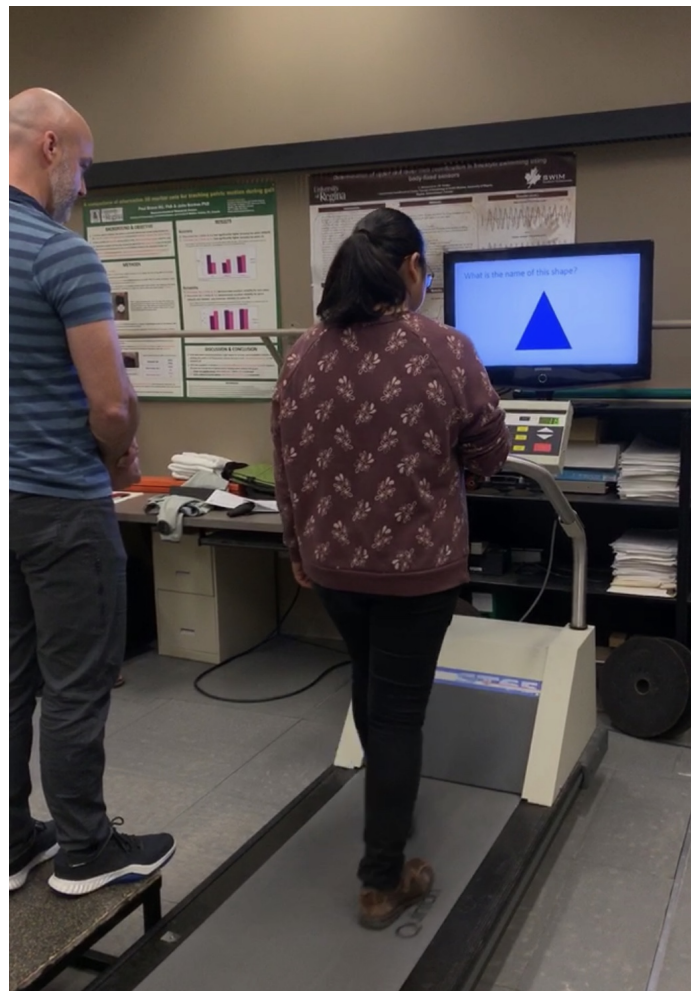

Figure 1

The physical setup of the controlled laboratory study at one of the data collection sites (photo taken during pilot testing, prior to discovering the need to require two-handed use).

Data Collection Procedures

The data collection in this study made use of two instruments. An online questionnaire was created to collect pre-study and post-study measures (e.g., demographics, walking pace), as well as pre-task and post-task measures (e.g., task familiarity, interest, and expected difficulty; perceptions of the degree and impact of the interruptions). We also asked participants three question in the post-task questionnaire to assess their level of task engagement, which were based on work by O'Brien and Toms (2010) and adapted by Harvey and Pointon (2017). All factual data was collected as precise answers or option selections; all perception data was collected using five-point Likert scales. Note that since participants were asked these questions while walking on the treadmill, the researchers showed the questions and possible responses on a tablet, read them to the participants, and collected responses on their behalf. The second data collection instrument was a bespoke iOS app running on the provided mobile device, which generates logs of eye tracking events and use of the search interface. More detail on this app is provided in the section that follows. A summary of the dependent variables and how they were collected is provided in Table 3.

Observational data was collected via field notes written at the conclusion of each participant session. While it may have been possible to take field notes in the midst of the search sessions, doing so may have been perceived by the participants as judging their behaviour, which we wished to avoid. Given that such post-study recording of observational data may not be particularly accurate and complete, we viewed this qualitative data as secondary data to corroborate and explain the primary quantitative data in this study.

The study was administered in two different research sites. Both followed a carefully-documented process, with the only differences being supplemental physical data collection that is beyond the scope of this paper (Site 1: motion sensor for gait analysis; Site 2: galvanic skin response (GSR) sensor for analysing stress levels). The common process was to: (1) greet the participant and collect informed consent; (2) get the participant set up with the physical data collection devices; (3) have the participant start walking on the treadmill at a comfortable pace; (4) collect the pre-study data; (5) show the participant the mobile device and give them training on its use; (6) show the participant the app and run through the app training and calibration procedure; (7) explain the study procedures and provide examples of interruption activities and tasks; (8) allow the participant to use the app on the training task so that we could verify that the interruption detection events were being captured correctly; (9) show the target task and administer the pre-task questionnaire; (10) allow the participant to complete the task, interrupt them as required by the interruption order specified in the Graeco-Latin square; (11) administer the post-task questionnaire; (12) repeat steps 9-11 for the second and third tasks; (13) administer the post-study questionnaire; (14) allow the participant to stop and step off the treadmill; (15) answer any questions the participant has and thank them for their time. Study sessions lasted between 30 and 45 minutes.

Participants were recruited from within the two research sites, with the primary criteria being that they were experienced mobile device users, use their mobile devices while walking, regularly conduct web searchers on their mobile devices, and were able to walk 20 minutes without pain or fatigue. These selection criteria were verified in the pre-study ques-

Variable	Data Collection Instrument	Data Type	Number of Questions
age	pre-study questionnaire	numerical	1
gender identity	pre-study questionnaire	option selection	1
level of education	pre-study questionnaire	option selection	1
smartphone use	pre-study questionnaire	Likert scale	4
comfort with walking	pre-study questionnaire	Likert scale	4
task familiarity, interest, and expected difficulty	pre-task questionnaire	Likert scale	3 per task
degree of interruption on query stage	post-task questionnaire	Likert scale	1 per task
impact of interruption on query stage	post-task questionnaire	Likert scale	1 per task
degree of interruption on search results stage	post-task questionnaire	Likert scale	1 per task
impact of interruption on search results stage	post-task questionnaire	Likert scale	1 per task
degree of interruption on document stage	post-task questionnaire	Likert scale	1 per task
impact of interruption on document stage	post-task questionnaire	Likert scale	1 per task
task engagement	post-task questionnaire	Likert scale	3 per task
walking pace during study	post-study questionnaire	numerical (m/s)	1
normal walking pace when not using mobile device	post-study questionnaire	numerical (m/s)	1
length of interruption	app logs	numerical (s)	3 per task
time to resume the task	app logs	numerical (s)	3 per task
total time to task completion	app logs	numerical (s)	1 per task
relevance of each selected document	app logs	relevance score	minimum 3 per task

Table 3

The dependent variables and data collection instruments used to measure these variables.

tionnaire: all participants stated that they were either extremely or moderately comfortable using a mobile phone; all but two participants stated that they use their smartphones while walking on a daily basis; and all participants use their smartphones to search the web at least once a week. We did not control for any other individual differences among the participants. The research was reviewed and approved by the Research Ethics Boards at both research sites, and participants were given a chance to win a gift card for a local coffee shop (50% chance, based on the total number of participants in the study). Making the compensation for participation based on chance, participants are not obligated to include it as taxable income. Data was collected from 18 participants at the first research site, and 9 participants at the second for a total of 27 participants. Assuming a medium effect size (d=0.5), a significance level of 5%, power of 0.8, and given the repeated-measures design, power analysis suggests a requirement for a total of 27 participants.

Participants from the first site were all students - either undergraduate or postgraduate - while those from the second site were a mixture of students and staff. The median age of all participants was 26; those from the first site were generally younger (median of 25 years against a median of 34 years at the second site); ages over both sites ranged from 18 to 54. Eleven participants were undergraduate students, 10 were Master's-level, and 6 were either doctoral students or staff already holding doctoral qualifications. We note that, although it would have been interesting to include age as a dependent variable in our analyses, this would require larger numbers of participants and a more even distribution in the ages thereof. We discuss the potential effect of age later in the discussion section.

This study was carefully designed to control and vary the independent variable and collect data on the dependent variables in a reliable manner. A detailed procedures manual was created and care was taken to show the tasks and interruption events in a repeatable and consistent manner at both research sites. All of the data collection instruments and the study procedures were debugged via multiple rounds of pilot testing. As a result of this effort, we have a very high degree of confidence in the validity of the study results. All data collection instruments are provided as supporting documents with this manuscript, in order to support replication of this study and others to follow similar procedures in their own follow-up research.

Mobile Search Eye-Tracking App

A key element in this research is the development of an eye-tracking app to determine when the participants are looking at the mobile device and when they are looking away. The use of eye-tracking for gaze analysis during search tasks has become popular, revealing interesting insights into where people look when they conduct searches (Bhattacharya & Gwizdka, 2019; Kim et al., 2017). For our research, we did not require such a fine level of accuracy, and therefore, did not require the use of external eye-tracking hardware; instead we collected coarse data (whether the participant is looking at the device or away from it). By integrating the eye-tracking into the app and using the built-in hardware on the mobile device, we were able to collect this data in an unobtrusive and naturalistic manner. The app is made available on GitHub[2] for other researchers to replicate this study or use and extend for their own specific purposes.

––––––––
[2]https://github.com/shaheedinc/eye-tracker-ios

The TrueDepth front-facing camera on recent generations of the Apple iPhone (since model X) projects 30,000 infrared dots and uses an infrared camera to create a depth map of whatever is in front of the device (Apple Inc., 2019). While this was originally developed to permit unlocking the device using facial recognition, Apple has recently added features to its augmented reality library in iOS (ARKit) to support face-tracking and eye-tracking in real time.

Determining whether one is looking at the device or not requires knowledge about the orientation of the device in relation to one's eyes. We created a short calibration session each time the app is started that asks the user to follow a red dot shown at eight locations around the perimeter of the display. Measurements are taken during this process, which then define the extent of the virtual display. To account for the variability in the orientation of the device while walking, a buffer was added to increase the size of this virtual display, the extent of which was determined via informal experimentation and pilot testing. Since this calibration is only done once per participant, we asked that the participants hold the device with two hands, and reminded them to realign the device if it appeared that they were changing the angle during the course of the study.

While using the app, events are captured and stored locally. For the eye tracking, these events are based on detecting changes of state (from looking at the device to looking away, and from looking away to looking at the device), although not exactly wherethe participant is looking on a per pixel or screen element level. All other interactions with the website loaded in the app (for this study, Google Web Search) are also logged. The only exception to this logging is that of individual keystrokes while typing, which do not generate individual events in iOS as a security precaution. When the app is closed, the event log is saved to a cloud-based NoSQL database, where it can then be retrieved for data analysis purposes.

Data Analysis

Each of the tasks were designed to be reasonably similar in complexity and familiarity, and each interruption at a given level (none, low, high) were designed to be at a similar cognitive complexity level. To start, we verify that this was indeed the case, after which we aggregate the data across all tasks. For the time measurements, mean values and standard errors of the mean are calculated and presented graphically. ANOVA is used to check for statistical significance between all measures and then this is further investigated in a pairwise fashion to evaluate the differences between the pairs of conditions with a focus on how (order and magnitude) they differ from each other, not simply whether or not they differ. Doing each pair-wise comparison directly minimizes the Type I errors that can occur when multiple comparisons are done simultaneously, eliminating the need to account for these through procedures such as Bonferroni correction. We also conducted Shapiro–Wilk tests to confirm the normality of the data, where appropriate. We also report effect sizes for significant ANOVAs and non-parametric tests (η^2 and r respectively) and perform post hoc power analysis for significant ANOVAs.

For the questionnaire measurements of perceptions and opinions, the frequency of responses at each Likert scale level are counted and presented graphically. Since the distances between these levels are not equal, the data may not be normally distributed. As a result, we use the non-parametric statistical Mann-Whitney test to confirm pair-wise statis-

tical significance. In cases where all pairs are consistent, we instead use the Kruskal-Wallis test over multiple conditions.

As the qualitative data collected via field notes of our observations during the study sessions were treated as secondary data, they were not explicitly analyzed. Instead, we consulted these notes after analyzing the quantitative data to add context and explanation of observed behaviours as needed.

Results

We now present the results in the following order: (1) the similarity between the tasks and between the interruptions at a given cognitive complexity level; (2) the effects of the experimental conditions on the amount of time participants were interrupted during each of the three stages of their search tasks and the impact on overall task completion times; (3) the participants' subjective impressions of the degree and impact of the interruption on each stage of the search tasks; and (4) the relevance of the clicked documents and accuracy of query input, grouped by experimental condition.

Task and Interruption Similarity

Task Similarity

Although the tasks were selected for their similarity, we collected pre-task questionnaire data to verify that participants had similar impressions of familiarity, interest, and expected difficulty. Table 2 shows the mean values of these measures. Task familiarity was generally low: between moderately familiar (3) and slightly familiar (2) for task 1 and between slightly familiar (2) and not familiar at all (1) for tasks 2 and 3. A pair-wise statistical analysis revealed significant differences between task 1 and both tasks 2 (W=207, $p < 0.005$, r=0.63) and task 3 (W=170, $p < 0.001$, r=0.80).

Task interest was also relatively low, with responses generally between moderately interested (3) and slightly interested (2). There was no statistically significant differences between the tasks on this measure (χ^2=3.65, df=2, $p = 0.162$). Prior task interest had no significant effect on any of the measures discussed later in this paper.

Expectations of task difficulty followed a reverse trend to that of familiarity. With task 1 generally found more familiar, it followed that the expectation of difficulty was lower (between slightly difficult (2) and not difficult at all (1)). The less familiar tasks 2 and 3 had responses between moderately difficult (3) and slightly difficult (2). A pair-wise statistical analysis revealed significant differences between task 1 and task 2 (W=501, $p < 0.05$, r=0.32) but not between task 1 and task 3 (W=432, $p = 0.22$) or between tasks 2 and 3 (W=302.5, $p = 0.26$).

We also calculated the mean time to complete the tasks under the no interruption condition (last column, Table 2), removing a single obvious outlier that occurred on task 2. While the participants who were assigned task 1 for the no interruption condition spent more time on this task than than the participants who were assigned the other two tasks for this condition, the differences were not statistically significant (F(2,19) = 0.410, $p = 0.670$). Note that the timing data is incomplete, which will be explained further in the following section.

While the specific group of participants in this study did express more familiarity with one task than the other, and expected this task to be less difficult, the differences in interest and time to task completion for the no interruption condition were not significant. As a result, we conclude that, although there are some differences in the participants' prior knowledge about the tasks, they are sufficiently similar to allow for the aggregation of the other data across the tasks. Doing so adds to the statistical power of all further analyses.

Interruption Similarity

For each of the two interruption levels (low and high), three different activities were presented to the participants (see Table 1). We analyzed these to determine if there were any unintended differences in participants' perceptions of the interruptions (degree and impact of the interruption). This analysis revealed no significant differences for each of the no interruption condition (degree: χ^2=3.26, df=2, $p = 0.196$; impact: χ^2=0.404, df=2, $p = 0.817$), low interruption condition (degree: χ^2=1.28, df=2, $p = 0.528$; impact: χ^2=0.065, df=2, $p = 0.968$), or high interruption condition (degree: χ^2=0.464, df=2, $p = 0.793$; impact: χ^2=0.557, df=2, $p = 0.757$). While our further analysis keeps the interruptions at each stage of the search process distinct, this result allows us to consider the interruptions at each level of cognitive complexity to be internally consistent.

Interruption Duration and Time to Task Completion

As one would expect, by introducing interruptions of different levels of cognitive complexity, the time spent attending to these activities has an impact on the time to task completion. Our bespoke iOS app allowed us to measure the amount of time that a participant's gaze was focused on the screen and how much time their gaze (and attention) was elsewhere. Out of 27 participants, we were able to obtain complete and reliable data from 21. The six failed data collection episodes were typically due to a participant wearing thick glasses and/or frequently moving and tilting the mobile device. While these were not evenly distributed in the Graeco-Latin square (they are at positions 3, 5, 5, 6, 6, and 8), aggregating the data over the tasks and interruption levels minimizes the effects of this missing data. In addition, for one of the no interruption conditions, a data value was removed as an outlier (the time to task completion was nine standard deviations greater than the mean value without that data point included). While this missing data has an impact on the statistical power (reducing it from 0.8 to 0.7 for the timing data), all of the statistical analyses below had clear outcomes, and for those where statistical significance was found, medium to large effect sizes.

Interruption Duration

To understand the impact of the experimental conditions in this study, we start by analyzing how long the participants were interrupted during each stage of the search task. Figure 2 illustrates the mean time each participant spent looking away from the device due to the interruptions. Clearly, the high interruption condition resulted in a significant amount of time spent looking away from the device compared to the low interruption event. This difference proved to be statistically significant (query stage: F(1,41)=61.7, $p < 1 \times 10^{-8}$,

$\eta^2 = 0.61$, power= 1.0; search results stage: F(1,41)=52.0, $p < 1 \times 10^{-8}$, $\eta^2 = 0.57$, power= 1.0; document stage: F(1,41)=39.9, $p < 1 \times 10^{-6}$, $\eta^2 = 0.50$, power= 1.0).

Considering the data across the stages, there was a very small difference in the low interruption condition (recognition of simple shapes). For the high interruption condition, there were some minor differences in the interruption duration due to the different types of information the participants were asked to recall (i.e. the number of people in the participant's family, number of letters in the participant's first name, last digit of the participant's phone number) and slight differences in the calculation to be performed. However, these differences were not statistically significant (low interruption: F(2,62)=0.486, $p = 0.618$; high interruption: F(2,62)=1.63, $p = 0.204$), confirming sufficient similarity among both the low interruption (minor re-allocation of attention and working memory) and high interruption (moderate re-allocation of attention and working memory) activity sets.

Time to Task Completion

When considering the total time to task completion, the impact of both the low and high interruption conditions are evident. These results are illustrated on the left side of Figure 3. While the difference between the no interruption and low interruption conditions was not statistically significant (F(1,40)=1.96, $p = 0.169$), the difference between the low interruption and high interruption conditions was (F(1,41)=6.33, $p < 0.05$, $\eta^2 = 0.14$, power=0.723). This illustrates how the routine activity of the low interruption condition had only a minor impact on the time to task completion, while the moderate amount of attention and working memory re-allocation in the high interruption condition had a noticeable impact.

To further assess the interruptions' impacts, we consider the time to task completion with the interruption duration removed (see the right side of Figure 3). While noting the mean time to task completion increases, pairwise analysis of the data did not reveal any
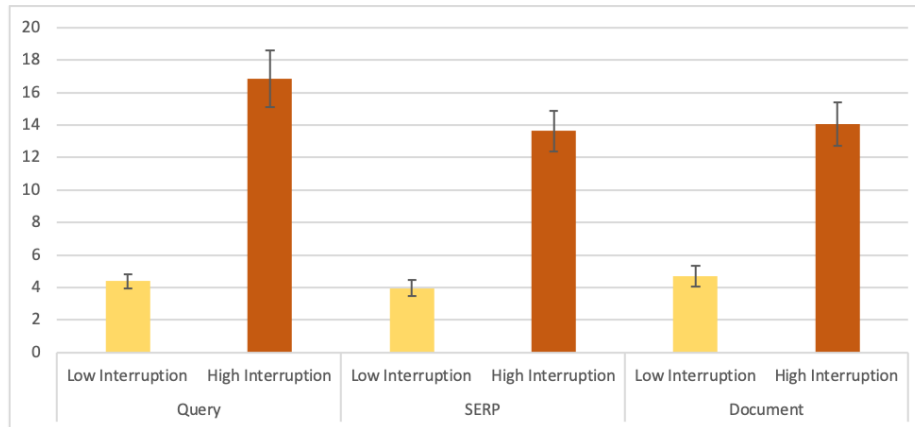

Figure 2

Mean time of the low and high interruption duration (seconds) during each stage of the aggregated search tasks. The error bars represent the standard error of the mean. Note that the no interruption condition is excluded here since the value would be zero seconds.

statistically significant differences (no interruption v. low interruption: $F_{(1,40)}=0.347$, $p=0.559$; low interruption v. high interruption: $F_{(1,41)}=0.461$, $p=0.501$; no interruption v. high interruption: $F_{(1,40)}=1.60$, $p=0.213$). These results suggest that, while there is a resumption lag that increases with the cognitive complexity of the interruption activity, it is not strongly influenced by the conditions in this study and may instead be a penalty that is influenced by other factors that were not controlled in this study (e.g., participants' short term memory).

While the average resumption lag was over 9 seconds between the no interruption and low interruption conditions, and over 20 seconds between the no interruption and high interruption conditions, these represent the total resumption lag for three different interruptions. Although we did not find differences between the experimental conditions, the effects of the interruptions should not be dismissed. This resumption lag represents the time the participants took to re-direct their attention back to the task at hand, reacquaint themselves with what they had been doing, and choosing what to do next.

Because our app captures both gaze events and other interaction events, we are able to determine how long it took participants to resume their tasks once they returned their attention to the device. Due to the aforementioned iOS security constraints, we cannot comment on the query stage. For the search results stage, the mean time to resume interacting was 1.5 seconds for the low interruption event, and 1.7 seconds for high interruption event. For the document evaluation stage, the mean time to resume interacting was 1.6 seconds for both the low and high interruption events. The differences were not statistically significant (search results: $F_{(1,41)}=0.091$, $p=0.764$; document evaluation: $F_{(1,41)}=0.001$, $p=0.971$). So while the participants were generally quick to resume interacting, the remainder of the resumption lag was represents the time taken to return back to the task. These findings illustrate the resiliency of the participants to recover from the interruptions and resume their search activities quickly.

We observed strategies that some participants employed to mitigate the interruptions and associated resumption lag. When entering the query, some participants quickly
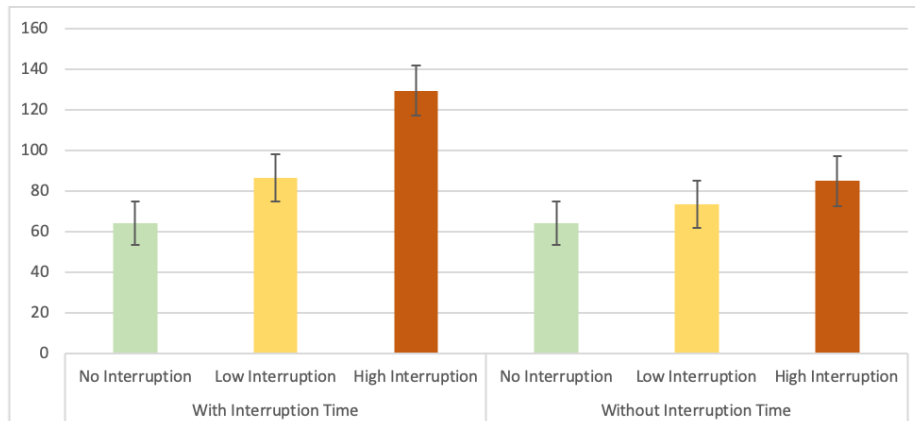

Figure 3

Mean time to task completion (seconds), with the duration of the interruption included (left) and excluded (right).

finished typing and submitted the query before attending to the interruption, even though they were explicitly instructed to focus on the interruption immediately when it occurred. When viewing the search results list, as an interruption event occurred, some participants would tap on whichever search result they were considering, and then attend to the interruption activity. In both of these strategies, the page loaded while they were looking away, enabling the participant to return to the task at a natural break in the activity (e.g., reviewing a freshly-loaded document). When an interruption occurred while viewing a document, some participants did not continue reading the document after the interruption was finished, and instead went back to the search results immediately. Together, this set of strategies allowed some participants to be more efficient in recovering from the interruptions. This is consistent with the literature, which suggests that interruptions that occur during course breakpoints in the cognitive activity are less damaging (Adamczyk & Bailey, 2004).

Subjective Impressions of the Interruptions

Degree and Impact of Interruptions

We now consider the participants' perceptions of the interruption events. For each stage of the search process (query, search result evaluation, and document evaluation), data was collected via post-task questionnaires regarding the degree of the interruptions and their impact on the primary tasks. For the no interruption condition, these questions were asked in relation to the ambient distraction of walking while searching, rather than the explicit interruptions.

The distributions of the responses for each interruption condition at each stage of the search process are illustrated in Figure 4. The mean values and statistical analyses are reported in Table 4. The left column shows that the participants perceived the interruptions at each level of cognitive complexity at an increasing degree (as we had intended), and that these differences were statistically significant. The right column shows that the participants perceived the impact of the interruptions to also be increasingly significant as the cognitive complexity of the interruption increased. The degree and impact of the interruption showed a similar pattern between the stages of the search process, which we confirmed to be not statistically significant.

Although we had previously shown that removing the time to attend to the interruption itself resulted in only a minor resumption lag, the perception of the degree and impact of the interruption on the search task were greater than this small resumption lag would suggest. The no interruption level was perceived as having the least degree of interruption and having the least impact; the low interruption level (minor re-allocation of attention and working memory) was perceived as having a moderate degree of interruption and a moderate impact; and the high interruption level (moderate re-allocation of attention and working memory) was perceived to have a high degree of interruption and a considerable impact on the search process.

Engagement in the Task

At the end of each task, we asked three questions regarding the degree of engagement in the task (forgetting immediate surroundings, losing track of time, being absorbed in the task). We expected that the no interruption condition would result in a higher

	Interrupt.	Degree	v. no interrupt.	v. low interrupt.	Impact	v. no interrupt.	v. low interrupt.
query	no	1.52			1.63		
	low	2.70	W=4.38, $p < 1x10^{-4}$		2.48	W=3.30, $p < 0.001$	
	high	3.67	W=5.70, $p < 1x10^{-7}$	W=3.21, $p < 0.005$	3.63	W=5.70, $p < 1x10^{-4}$	W=3.21, $p < 0.001$
results	no	1.67			1.74		
	low	2.56	W=3.60, $p < 0.001$		2.48	W=2.51, $p < 0.05$	
	high	3.48	W=5.14, $p < 1x10^{-6}$	W=3.19, $p < 0.005$	3.37	W=4.58, $p < 1x10^{-5}$	W=2.63, $p < 0.01$
doc.	no	2.00			1.70		
	low	2.48	W=2.21, $p < 0.05$		2.41	W=2.92, $p < 0.005$	
	high	3.56	W=4.45, $p < 1x10^{-5}$	W=3.87, $p < 0.001$	3.48	W=4.96, $p < 1x10^{-6}$	W=3.45, $p < 0.001$

Table 4

Perceived degree and impact of the interruption for each of the search stages (i.e., query, search results, and document) over the three experimental conditions (interruption levels). Results of Mann-Whitney-Wilcoxon tests are provided comparing to the no interruption (baseline) and the low interruption conditions. All comparisons are statistically significant.

degree of engagement, and that interrupting the participants would result in them being less engaged. The data did not confirm this hypothesis. The responses were consistent between all experimental conditions, generally between neither agreeing or disagreeing with the statements (3) and somewhat agreeing with the statements (2). For forgetting one's immediate surroundings, the mean responses for the no, low, and high interruption experimental conditions were 2.30, 2.52, and 2.60 (χ^2=1.01, df=2, $p = 0.604$). For losing track of time, the mean responses were 2.59, 2.56, and 2.56 (χ^2=1.00, df=2, $p = 0.951$). For being absorbed in the task, the mean responses were 2.26, 2.30, and 2.44 (χ^2=0.508, df=2, $p = 0.776$). It appears that the laboratory setting and the relatively low interest in the tasks kept participants from becoming deeply engaged in the activity.
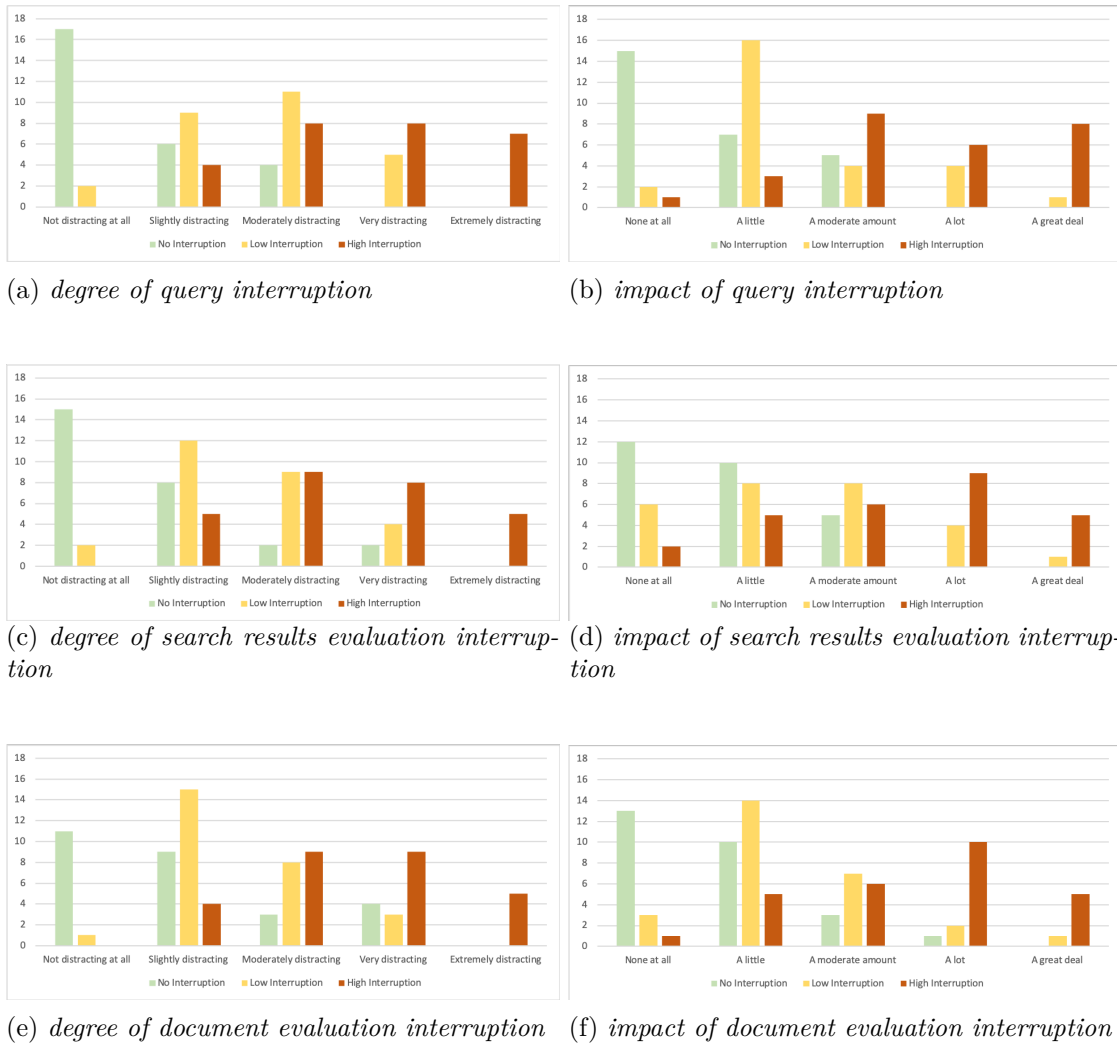

(a) *degree of query interruption*

(b) *impact of query interruption*

(c) *degree of search results evaluation interruption*

(d) *impact of search results evaluation interruption*

(e) *degree of document evaluation interruption*

(f) *impact of document evaluation interruption*

Figure 4

Frequency of responses to questions regarding the perceived degree and impact of the interruptions at each stage of the search process.

Query Input Accuracy and Document Relevance

We investigated whether the conditions had any impact on the participants in terms of being able to input the queries and select relevant documents. To do so, two experienced IR researchers manually annotated the query and click data obtained from the experiments. We assessed whether an inputted query had any spelling mistakes and, for each unique clicked document, whether it was relevant, partially-relevant (i.e., about the topic in general but not specific) or non-relevant. There was a high level of agreement between the two assessors for both the queries ($\kappa = 0.803, z = 6.24, p \ll 0.01$) and the document relevance judgements ($\kappa = 0.753, z = 5.83, p \ll 0.01$). Any disagreements were resolved between the assessors, resulting in a set of final judgements.

Although one might expect otherwise, it seems that experimental conditions had no significant effect on query input accuracy ($\chi^2 = 2.43, df = 2, p = 0.297$), although there were relatively fewer errors made under the no interruption condition. Only 2 of the uninterrupted queries contained errors (9%), while 27% of the low interruption and 22% of the high interruption queries had errors. Note that the participants were all experienced mobile device users who regularly use them while walking, which could account for their skill in typing. Also note that we instructed the participants to not use query auto-completion and to type the queries fully themselves. Although our raw results would suggest that there were more errors in the low interruption condition than the high interruption condition, there was no significant difference between the two conditions for this measure, and so we attribute this to individual differences among the participants.

Likewise, when considering clicked document relevance there is no significant difference between the three conditions ($\chi^2 = 3.37, df = 4, p = 0.498$). The majority of documents (104) were judged to be relevant, while 32 were partially-relevant and only 2 were non-relevant. The partially-relevant and non-relevant documents were distributed somewhat randomly among the experimental conditions. As such, we conclude that the interruptions themselves did not have an impact on the participants' abilities to make good relevance judgements, despite increasing task time and the effects on their perceptions.

Discussion

Key Contributions

The work presented here makes two key contributions to the existing literature. The first of these is the methodology developed to induce simulated interruptions in a controlled laboratory environment, and the iPhone app developed to log search activities and detect when attention has been diverted away from and back to the device. This combination of methodology and technology is novel, providing a reliable and repeatable mechanism for simulating interruptions in mobile search tasks that are at different levels of cognitive complexity, and measuring the associated impacts on the search process.

The second contribution is that this is the first user study in which repeatable simulated interruptions have been used to demonstrate how such events affect mobile search performance and searcher's perceptions of the degree and impact of the interruptions. Our findings augment the existing literature on the cognitive processes involved in dealing with an interruption and subsequently returning to the primary task. This approach differs from

previous research in the field by inducing specific interruptions under controlled experimental conditions while participants are performing predefined search tasks. It builds on existing work on SIIDs by extending these concepts to study their effects on multi-step tasks (i.e., web search).

RQ1 - How do interruptions of differing levels of complexity affect mobile search task performance?

We confirmed that the greater the complexity of the interruption, the longer the duration of the interruption. This resulted in significant differences in the time to task completion. Interestingly, when removing the duration of the interruptions themselves, the pattern of the penalty of being interrupted as measured in the resumption lag increased with the cognitive complexity of the interruption, but not at a statistically significant level. While this means that we cannot claim that this time penalty is influenced by the complexity of the interruption, it still represented an average of over 3 seconds per interruption at the low level of complexity (minor attention and working memory re-allocation) and over 6 seconds in the high level of complexity (moderate attention and working memory re-allocation). As such, when one's attention is fragmented due to an interruption, there is a non-trivial resumption lag that will occur, regardless of how cognitively complex the interruption is.

The fact that the resumption lag is not higher may be due to participants engaging in specific mitigating strategies, which we observed in several instances. For example, when interrupted in the process of assessing documents in a search result list, some participants tapped on the link to whatever document they were considering at the time of the interruption. This meant that the mobile device was loading this document while they were attending to the interruption event, allowing them to resume their search tasks efficiently. This is similar behaviour to the *compensatory strategies* employed by people when their cognitive resources are contended (Ericsson & Kintsch, 1995; Oulasvirta et al., 2005). Qualitative studies of these strategies are needed to reveal the conditions in which they are employed, and their potential for negatively impacting the outcomes of the search tasks.

It is also worth noting that since the measurements were made over a subset of the information seeking process (post-query formulation), this represents a lower bound on the objective impact of the interruptions. Further research on the impact during query formulation could reveal interesting effects, especially given the cognitively complex nature of query formulation and how it is intertwined with the interaction of typing on the mobile device. This may also explain why our findings in terms of interruptions causing typographical/input errors do not entirely align with those of Schaller et al. (2012). Their findings are based on logs from an "in-the-wild" study, where interruptions could not be controlled at all and where users were navigating around a busy city centre during a festival, often whilst also interacting with other attendees/friends. The environment in our study is considerably more controlled and, although participants were interrupted during query input, this was done just once while the participants entered a given query.

RQ2 - How do interruptions of differing levels of cognitive complexity affect the subjective impressions of the impact of the interruptions and engagement in the search task?

We confirmed that the greater the cognitive complexity of the interruption, the greater the perceived degree of the interruption and perceived impact on the search task. However, it did not affect the participants' degree of engagement in the task. The first finding is consistent with that of earlier work on distractions and mobile search (Harvey & Pointon, 2017, 2019); however, the lack of impact on engagement differs from this prior research, suggesting that different levels of atomic interruptions impact engagement differently than different levels of constant movement (i.e., walking on a treadmill instead of sitting at a desk). The lack of impact on engagement in the task may also have been influenced by the relatively low interest the participants showed in the tasks; studies using personally important tasks may reveal different results when interruptions are induced in the midst of the search process.

Implications

Even though participants showed great resiliency in recovering from interruptions regardless of their cognitive complexity, there remained a non-trivial resumption lag that was present for both the low and high interruption conditions. The participants' impressions of the degree of the interruptions and their impact on the search activity aligned with the different levels of cognitive complexity of the interruption. Together, these findings show that even though the objective impact of the interruptions are not affected by their cognitive complexity, the subjective perceptions of them are. While some participants employed compensatory strategies to create natural breaks in the search activity when interruptions occurred (e.g., hastily tapping on whichever search result they were considering so that it could load while the searcher attended to the interruption), doing so is not an effective information retrieval strategy. We contend that more research is needed on these strategies, the factors that contribute to the resumption lag, and the factors that contribute to the searcher's impressions of the impact of the interruptions. There is also a need for more work on the design and study of mobile search interfaces that provide mechanisms to encourage and support strategies for the resumption of search tasks after interruptions occur, perhaps mitigating the negative impressions of the impacts of the interruptions.

To this point, we propose the following ideas for mobile search interface design. We have already illustrated in this research how the front-facing camera on a mobile device can be used to detect when a searcher's gaze leaves the device. Within this same framework, we could detect what was the last thing the searcher viewed before looking away, and whether the gaze away from the device is sufficiently long to infer that an interruption has occurred. If this is the case, the last viewed element may be highlighted using visualization methods (e.g., encapsulated in a bounding box or highlighted with a subtle background colour) to allow the searcher to easily identify where they had been last looking, minimizing the resumption lag and allowing the searcher to readily resume their task. Alternately, if there is no interaction with the device for sufficiently long period of time (e.g., about 4 seconds for the low interruption condition, and 13 seconds for the high interruption condition, as shown in Figure 2), one might also infer that an interruption has occurred and highlight

the last element interacted with on the interface. The sensors on the device may also be used to detect interruptions (e.g., the device is moved in a motion that implies that the user cannot see it anymore, such as moving it down to one's side). The specific methods used, and their mechanisms for helping the searcher to resume the task must be done in consideration of the search stage. Such approaches may further enhance the searchers' resiliency in recovering from interruptions, minimizing the negative subjective impression of the impact of the interruption.

Generalizability and Limitations

Although this study was conducted in a controlled laboratory setting, some of the findings may be generalized beyond the lab. While the participants walked on a treadmill, care was taken to ensure that the pace was realistic for each individual participant; participants who chose slow speeds were asked whether this truly reflected their natural everyday walking pace and adjustments were made accordingly. The interruptions were designed to mimic attention and working memory re-allocation that occur in real-world interruptions. For example, the *high interruption* condition is analogous to being stopped in the street mid-search by someone asking for directions or for a donation. Although the interruptions themselves were explicitly designed to deviate from realism so that we could induce a somewhat consistent re-allocation of attention and working memory, their basis in cognitive psychology suggest that findings can be generalized to interruptions that have similar cognitive effects. While the lack of authenticity of the interruptions is a limitation of the study, we expect to find similar results for naturalistic studies that produce interruptions at a similar level of cognitive complexity, which we will confirm in future work.

Where we believe things may not generalize is in terms of the search tasks and process. While the tasks were chosen to be of the type that are common in mobile search settings (e.g., fact-finding), the topics themselves were not of particular interest to the participants. While this was by design (to avoid the risk of participants having conducted similar searches recently), it does introduce a limitation to the study. In particular, it resulted in a relatively low level of interest and engagement in the tasks; it is unclear how things would differ for tasks in which participants are personally invested. In addition, this study focused on a subset of the information seeking process, starting at query input. Given that query formulation may incur a significant cognitive load (Gwizdka, 2010), the impact of interruptions at this stage may be greater than what was observed in this study. These represent interesting avenues for future work.

We acknowledge the differences in participant ages, although we could not find any effects of age on any of the measurements. This is something that could have an observable effect with a larger sample, particularly with older users or those with less mobile search experience; however, this was not the aim of our study. Finally, we note that, due to some data collection being incomplete/unusable, the sample we used for our analyses of the duration of the interruptions or the time to task completion was not as large as we would have liked and does not guarantee a statistical power of 0.8. This may have resulted in Type II errors, and some of our negative results may in fact turn out to be valid hypotheses with a larger number of participants. However, our positive findings are based predominantly on tests that achieved medium to large effect sizes.

Conclusions & Future Work

Overall, while we found that increasing the cognitive complexity of the interruption had minimal impact on task performance once the duration of the interruption itself was removed, the participants' perceptions of the impact of the interruption increased with the complexity. That is, the subjective impacts on the mobile web search activities were greater than the objective impacts. We found evidence of behaviours that searchers undertake to mitigate the effects of the interruptions, minimizing the objectively measurable impacts of the interruption. Even so, the fact that the participants perceived the interruptions to be distracting and to have an impact on their performance tells us that we need to do more to support searchers when they are interrupted, and to learn more about the interplay between interruptions and mobile search behaviour.

Our future work will focus on six streams of research: (1) extending this specific study to determine the effect of the interruptions on movement and stress levels; (2) conducting a qualitative study investigating what kinds of interruptions occur in real-life, how people react to them, and the strategies they use to overcome them; (3) replicating these studies with tasks that are more difficult and of personal interest to participants, including tasks that are more typical of those performed "on the go" such as spontaneous, internally-induced, and in-context tasks like searching for the directions for somewhere to go for lunch, and by introducing a time constraint for the tasks; (4) conducting further controlled laboratory studies to examine the degree to which the interruptions induce the intended re-allocation of attention and working memory, the effects of the interruption timing within the search task stages, and the effects of interruptions on participants of differing cognitive abilities; (5) conducting similar studies in naturalistic settings to gauge the prevalence of different types of interruptions; and (6) designing and developing mobile search interfaces that mitigate the effects of interruptions by dynamically adapting to the searcher's attentional context. We note that (3) could also be developed further by considering more complex fact-finding tasks that would require more than a single query to satisfy. This would permit the effects of interruptions to be investigated at a whole session level, rather than just the effects on single search queries.

Acknowledgements

References

Adamczyk, P. D., & Bailey, B. P. (2004). If not now, when?: The effects of interruption at different moments within task execution. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 271–278. https://doi.org/10.1145/985692. 985727

Aliannejadi, M., Harvey, M., Costa, L., Pointon, M., & Crestani, F. (2019). Understanding mobile search task relevance and user behaviour in context. *Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval*, 143–151. https://doi.org/10.1145/3295750.3298923

Apple Inc. (2019). *About Face ID advanced technology.* https://support.apple.com/en-ca/HT208108

Bâce, M., Staal, S., & Bulling, A. (2020). Quantification of users' visual attention during everyday mobile device interactions. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Barnard, L., Yi, J. S., Jacko, J. A., & Sears, A. (2007). Capturing the effects of context on human performance in mobile computing systems. *Personal and Ubiquitous Computing*, *11*(2), 81–96. https://doi.org/10.1007/s00779-006-0063-x

Bergstrom-Lehtovirta, J., Oulasvirta, A., & Brewster, S. (2011). The effects of walking speed on target acquisition on a touchscreen interface. *Proceedings of the International Conference on Human Computer Interaction with Mobile Devices and Services*, 143–146. https://doi.org/10.1145/2037373.2037396

Bhattacharya, N., & Gwizdka, J. (2019). Measuring learning during search: Differences in interactions, eye-gaze, and semantic similarity to expert knowledge. *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, 63–71. https://doi.org/10.1145/3295750.3298926

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

Caird, J. K., Willness, C. R., Steel, P., & Scialfa, C. (2008). A meta-analysis of the effects of cell phones on driver performance. *Accident Analysis & Prevention*, *40*(4), 1282–1293. https://doi.org/10.1016/j.aap.2008.01.009

Carrascal, J. P., & Church, K. (2015). An in-situ study of mobile app & mobile search interactions. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2739–2748. https://doi.org/10.1145/2702123.2702486

Church, K., & Oliver, N. (2011). Understanding mobile web and mobile search use in today's dynamic mobile landscape. *Proceedings of the International Conference on Human Computer Interaction with Mobile Devices and Services*, 67–76. https://doi.org/10.1145/2037373.2037385

Church, K., & Smyth, B. (2009). Understanding the intent behind mobile information needs. *Proceedings of the 14th International Conference on Intelligent User Interfaces*, 247–256. https://doi.org/10.1145/1502650.1502686

Church, K., Smyth, B., Bradley, K., & Cotter, P. (2008). A large scale study of European mobile search behaviour. *Proceedings of the International Conference on Human Computer Interaction with Mobile Devices and Services*, 13–22. https://doi.org/10.1145/1409240.1409243

Couffe, C., & Michael, G. A. (2017). Failures due to interruptions or distractions: A review and a new framework. *American Journal of Psychology*, *130*(2), 163–181.

Crestani, F., Mizzaro, S., & Scagnetto, I. (2017). *Mobile information retrieval.* Springer Nature. https://doi.org/10.1007/978-3-319-60777-1

Dischler, J. (2015). *Building for the next moment.* https://adwords.googleblog.com/2015/05/building-for-next-moment.html

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211.

Gomes, S., & Hoeber, O. (2021). Supporting cross-session cross-device search in an academic digital library. *Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval*, 337–341. https://doi.org/10.1145/3406522.3446055

Gopi, R., & Hoeber, O. (2016). TwIST: A mobile approach for searching and exploring within Twitter. *Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval*, 273–276. https://doi.org/10.1145/2854946.2854991

Gwizdka, J. (2010). Distribution of cognitive load in web search. *Journal of the American Society for Information Science and Technology*, *61*(11), 2167–2187. https://doi.org/10.1002/asi.21385

Harvey, M., & Pointon, M. (2017). Searching on the go: The effects of fragmented attention on mobile web search tasks. *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 155–164. https://doi.org/10.1145/3077136.3080770

Harvey, M., & Pointon, M. (2019). Understanding in-context interaction: An investigation into on-the-go mobile search. *Information Processing & Management*, *56*(6). https://doi.org/10.1016/j.ipm.2019.102089

Hoggan, E., Crossan, A., Brewster, S. A., & Kaaresoja, T. (2009). Audio or tactile feedback: Which modality when? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2253–2256. https://doi.org/10.1145/1518701.1519045

Kamvar, M., & Baluja, S. (2006). A large scale study of wireless search behavior: Google mobile search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 701–709. https://doi.org/10.1145/1124772.1124877

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, *3*(1–2), 1–224. https://doi.org/10.1561/1500000012

Kim, J., Thomas, P., Sankaranarayana, R., Gedeon, T., & Yoon, H.-J. (2017). What snippet size is needed in mobile web search? *Proceedings of the ACM SIGIR Conference on Conference Human Information Interaction and Retrieval*, 97–106. https://doi.org/10.1145/3020165.3020173

Lamberg, E. M., & Muratori, L. M. (2012). Cell phones change the way we walk. *Gait & Posture*, *35*(4), 688–690.

Larsen, H. H., Scheel, A. N., Bogers, T., & Larsen, B. (2020). Hands-free but not eyes-free: A usability evaluation of siri while driving. *Proceedings of the ACM SIGIR Conference on Conference Human Information Interaction and Retrieval*, 63–72. https://doi.org/10.1145/3343413.3377962

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences*, *9*(2), 75–82.

Lin, M., Goldman, R., Price, K. J., Sears, A., & Jacko, J. (2007). How do people tap when walking? an empirical investigation of nomadic data entry. *International Journal of Human-Computer Studies*, *65*(9), 759–769.

Mayer, S., Lischke, L., Woźniak, P. W., & Henze, N. (2018). Evaluating the disruptiveness of mobile interactions: A mixed-method approach. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3173980

Mizobuchi, S., Chignell, M., & Newton, D. (2005). Mobile text entry: Relationship between walking speed and text input task difficulty. *Proceedings of the International Conference on Human Computer Interaction with Mobile Devices and Services*, 122–128. https://doi.org/10.1145/1085777.1085798

Morris, D., Ringel Morris, M., & Venolia, G. (2008). SearchBar: A search-centric web history for task resumption and information re-finding. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1207–1216. https://doi.org/10.1145/1357054.1357242

Nicolau, H., & Jorge, J. (2012). Touch typing using thumbs: Understanding the effect of mobility and hand posture. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2683–2686. https://doi.org/10.1145/2207676.2208661

O'Brien, H. L., & Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, *61*(1), 50–69.

Oulasvirta, A., Tamminen, S., Roto, V., & Kuorelahti, J. (2005). Interaction in 4-second bursts: The fragmented nature of attentional resources in mobile hci. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 919–928. https://doi.org/10.1145/1054972.1055101

Oviedo-Trespalacios, O., Haque, M. M., King, M., & Washington, S. (2016). Understanding the impacts of mobile phone distraction on driving performance: A systematic review. *Transportation Research Part C: Emerging Technologies*, *72*, 360–380.

Sarsenbayeva, Z., van Berkel, N., Velloso, E., Kostakos, V., & Goncalves, J. (2018). Effect of distinct ambient noise types on mobile interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(2), 1–23. https://doi.org/10.1145/3214285

Schaller, R., Harvey, M., & Elsweiler, D. (2012). Entertainment on the go: Finding things to do and see while visiting distributed events. *Proceedings of the Information Interaction in Context Conference*, 90–99. https://doi.org/10.1145/2362724.2362743

Sohn, T., Li, K. A., Griswold, W. G., & Hollan, J. D. (2008). A diary study of mobile information needs. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 433–442. https://doi.org/10.1145/1357054.1357125

Teevan, J., Karlson, A., Amini, S., Brush, A., & Krumm, J. (2011). Understanding the importance of location, time, and people in mobile local search behavior. *Proceedings of the International Conference on Human Computer Interaction with Mobile Devices and Services*, 77–80. https://doi.org/10.1145/2037373.2037386

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 449–455. https://doi.org/10.1518/001872008X288394

Wobbrock, J. O. (2019). Situationally-induced impairments and disabilities. In Y. Yesilada & S. Harper (Eds.), *Web accessibility* (pp. 59–92). Springer. https://doi.org/10.1007/978-1-4471-7440-0_5

Wolf, F., Kuber, R., Pawluk, D., & Turnage, B. (2017). Towards supporting individuals with situational impairments in inhospitable environments. *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility*, 349–350. https://doi.org/10.1145/3132525.3134783

Yesilada, Y., Harper, S., Chen, T., & Trewin, S. (2010). Small-device users situationally impaired by input. *Computers in Human Behavior*, *26*(3), 427–435. https://doi.org/10.1016/j.chb.2009.12.001