Examining field usability testing outcomes conducted on virtual and physical mobile devices using wearable eye tracking

by

Timothy McGowan

May 2019

Presented to the
Division of Yale Gordon College of Arts and Sciences
University of Baltimore

In Partial Fulfillment
of the Requirements for the Degree of
Master of Science

Approved by: ________________________________

[Kathryn Summers, Thesis Advisor]

____

[Greg Walsh, Committee Member]

Abstract

There is a growing need to conduct usability testing with smartphones and tablets as adoption of these devices continues to increase. Virtual device testing is a popular approach due to the diversity of devices and operating systems available to users in the mobile market; however, this approach can distance the results from real-world applications. Using a wearable eye tracker, this study compares how the use of a desktop emulator or physical device influences usability testing in a field environment. Results suggest that when compared to a virtual device (emulator), the use of a physical device has an influence on both time on task and perceived usability. Analysis of eye tracking measures identified differences in the amount of time that attention was given to specific areas on-screen and off-screen, indicating that using an actual device rather than a virtual device can also influence attention.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Mobile devices, such as smartphones and tablets, provide users with the flexibility to access the internet and popular applications at any time or location. Adoption of these devices has steadily increased across all markets and user groups, in pace with regular improvements to hardware and an expanding list of available features. Results from a 2018 survey conducted by Pew Research Center show that approximately 77% of all Americans now own smartphones, and approximately 1 in 5 now access the internet solely via their mobile device. Their portability and utility make them a valuable tool in many aspects of everyday life, and their continuous presence encourages strong feelings of attachment (Meschtscherjakov, Wilfinger, & Tscheligi, 2014). More and more users now reach for their mobile devices to complete complex tasks, even when desktop computers are readily available.

For designers and developers, the adoption of mobile devices has led to challenges in designing digital content. Compared to desktops and laptops, the mobile environment is subject to constraints that include display size and resolution, device hardware, and input mechanisms. The diverse market of devices and operating systems available to users requires extensive functional and usability testing. Available on-screen space, or viewport, is significantly restricted on a mobile device, impacting users' abilities to interpret detailed information within the context of the larger website. Designs also need to account for the increased potential for distraction when using mobile devices, including device notifications and external interruptions (Anderson et al., 2018; Duh, Tan, & Chen, 2006; Meißner, Pfeiffer, Pfeiffer, & Oppewal, 2017; Steil, Müller, Sugano, & Bulling, 2018).

Eye tracking has become a popular approach for conducting usability evaluations on various interfaces, for detecting and understanding usability problems and measuring the attraction and performance of design elements. In the area of mobile design, eye tracking research has been conducted to evaluate the usability of design prototypes

(Borys & Milosz, 2015; Kuhnel, Seiler, Honal, & Ifenthaler, 2017) and to determine how users' attention and reading behavior change by device and screen size (Biedert, Dengel, Buscher, & Vartan, 2012; Botella, Moreno, & Peñalver, 2014; Cuadrat Seix, Veloso, & Soler, 2012; Kim, Thomas, Sankaranarayana, Gedeon, & Yoon, 2015). However, studies like these have been limited to the use of desktop emulation of mobile devices and other novel setups, only available in controlled environments, in order to capture accurate date.

Results from non-eye-tracking usability studies suggest that the lack of field testing and the use of desktop emulators can influence the outcomes of usability testing of mobile applications (Coursaris & Kim, 2011; Duh et al., 2006; Levulis & Harris, 2015). While prior eye tracking studies have been limited to the use of fixed devices in controlled environments, the accuracy and precision of wearable eye trackers have improved to the point where it is possible to capture gaze data of a mobile device and the users' visual surroundings. This study will compare how the use of a desktop emulator or physical device influences usability testing outcomes in a field environment. Through a series of goal-based tasks, the use of the wearable eye tracker will allow for tracking of gaze data both on-screen and off-screen, increasing understanding of how mobile devices influence user attention.

Chapter 2: Theoretical Background

Eye tracking in usability research

Usability testing has often focused on performance metrics such as time-on-task, task success, and user satisfaction to compare design effectiveness (Lazar, Feng, & Hochheiser, 2010); however, it can be difficult to ascertain user engagement from these measures alone. Eye-tracking studies have been employed in a variety of research fields and subject domains to determine users' visual attention in structured testing environments and real-world observations. Within the field of Human Computer Interaction (HCI), eye tracking has become a popular tool in user testing for detecting and understanding usability problems, and for measuring the attraction to and performance of users interacting with stimuli in the design (Bojko, 2013; Borys, Czwórnóg, & Ratajczyk, 2016). Eye tracking provides an opportunity to obtain insights into the users' cognitive processes that take place between easily observable events, such as mouse clicks or touch gestures, and provide reasonable measurements of how users utilize attentional capacity in respect to available visual stimuli.

Eye tracking technologies are divided into two types based on their placement in relation to the subject and observed stimulus. Device-mounted trackers, also referred to as remote eye trackers, are mounted to a fixed stimulus and provide several benefits to research studies that measure user interactions within a constrained viewing area (Bojko, 2013). They offer a lower level of intrusion to participants, ensure a measure of recording consistency between individual participant sessions, and allow for more reliable automated analysis of gaze data. However, these types of devices need to be placed in a consistent environment to ensure comparable data collection between sessions, data collection is limited to the viewing area of the stimulus, and the fixed distance between the eye-tracker and participant needs to be maintained. Even minor shifts in the position of the participant or stimulus can impact the accuracy of reported gaze data.

Wearable trackers allow for tracking of participants’ gaze within the context of real-world environments. As the name suggests, the device is worn on the participant, often as a pair of glasses, and unlike with device-mounted trackers the distance between eye-camera and the participant is fixed. A forward-facing camera captures the entire field of view, allowing for dynamic interactions with the stimulus and a full range of movement in the environment. In contrast to remote eye trackers, wearable trackers are more obtrusive and participants’ awareness of the device may impact their natural behavior in the environment (Magnussen, Zachariassen, Kharlamov, & Larsen, 2017). Accuracy is limited by the resolution of the forward-facing camera and by changes in lighting conditions. Further, the dynamic relationship between participant and stimulus when using wearable eye trackers requires much of the data to be mapped manually (Benjamins, Hessels, & Hooge, 2018). In practice, use of wearable eye trackers has been directed towards observing behaviors and interactions with physical environments, such as evaluating the design and layout of a retail shopping experience (Harwood & Jones, 2014). However, recent improvements in capturing hardware and image analysis have made it more feasible to conduct mobile usability studies and other research applications using wearable eye tracking systems (Khamis, Eiband, Zürn, & Hussmann, 2018).

Eye tracking with mobile devices

Eye tracking has been incorporated into a number of studies to examine user behaviors and conduct usability testing with mobile devices and mobile device emulators. Studies have been predominantly task-based, are conducted in a laboratory, and often incorporate some form of secondary research measure including post-task surveys (Kuhnel et al., 2017), web analytics (Borys, Czwórnóg, & Ratajczyk, 2016), or device-sensor data (Biedert et al., 2012). Currently, device-mounted eye tracking is the most common approach, using a desktop emulator or a mobile device that is fixed in-place, as this approach reduces setup time and analysis (Cheng, 2011). However, there has been a growing number of studies that have applied wearable tracking in the past few years with the goal of better approximating a real-world testing experience.

Research with mobile devices and tablets have noted changes in user behavior and performance based on the size of the current viewport, suggesting that screen size can have an influence on the user experience (Botella et al., 2014; Kim et al., 2015). Desktop-mounted trackers have been used in several studies to compare changes to attention at various screen resolutions. For example, Cuadrat, Veloso, and Soler observed that users exhibit less eye movement on smaller screens in a visual search task (2012). Search engine behaviors have also been observed to be affected by screen size, with users focusing more attention on the center and top half of the screen (Fichter & Wisniewski, 2016; Lagun, Hsieh, Webster, & Navalpakkam, 2014), and influencing the depth of the search (Kim et al., 2015). Using a device-mounted eye tracker, Biedert et al. identified three types of scrolling strategies (entire page, line-by-line, and block scrolling) that users use interchangeably depending on the alignment of text on a smartphone (2012). Similar to the goals of this study, Levuis and Harris evaluated the use of emulators and physical devices in usability testing, noting small differences in performance, attention, and satisfaction (Levulis & Harris, 2015). However, the tests were conducted in a controlled environment and the physical device was fixed to the tracker which doesn’t allow for a more natural, hand-held device use.

Unlike device-mounted trackers, wearable eye tracking has been applied to mobile device studies in a variety of settings. Tupikovskaja-Omovie and Tyler applied wearable tracking to allow users to more naturally handle a mobile device in the lab, and used the subsequent recordings to develop journey maps of an online mobile shopping experience (2018). Other studies have designed the laboratory environment to approximate an office working experience with a tablet device (Borys & Milosz, 2015) and examined the use of wearable eye tracking in combination with a virtual environment to improve automated tracking of areas of interest (Meißner et al., 2017). In a recent field study, Ohm, Müller, & Ludwig demonstrated the ability to track attention on and off of a tablet device while navigating an indoor environment (2017). Results from these prior studies suggest that a comparison of virtual and physical mobile devices is possible using a wearable eye tracker in a field study.

Visual attention & distraction in task-oriented internet use

Attention is a highly focused and selective cognitive process which incorporates incoming sensory information with components of perception and memory. While there are multiple working theories on attention, it is generally accepted that attention is limited in capacity, takes effort to sustain, and is influenced by a combination of environmental stimulus, motivations, emotions, and memory (Nogueira & Ferreira, 2016; Tamber-Rosenau & Marois, 2016). Within the field of human computer interaction, there is a focus on specific components of visual attention related to top-down and bottom-up influences (Anderson et al., 2018; Bojko, 2013; Harwood & Jones, 2014; Johnson, 2014). Top-down describes the tendency for attention to be drawn to specific features based on prior knowledge or experience, and in many cases relates to an intention to complete a task or goal. In contrast, bottom-up describes the more reflexive influence on attention of low-level salient features including motion, luminance, and density.

Eye tracking has been used to examine the relationship between top-down and bottom-up influences in a variety of task-based studies. Research within the areas of video game design and web design has shown that prior experience and design expectations have an influence in directing search patterns and interaction behaviors (Almeida, Veloso, Roque, & Mealha, 2011; Pearson & van Schaik, 2003). In a real-world navigation activity, both task instructions and design elements of a mobile app have been shown to direct attention to specific features in the physical environment (Foulsham & Kingstone, 2012). Experiments in gaze control have shown that the ability to ignore distractions changes with age and that, in the absence of a task, low level features become more salient (Holmberg, Sandberg, & Holmqvist, 2014). Low level features have also been shown to draw attention away based on their relevance to current motivations and proximity to relevant features (Dumais, Buscher, & Cutrell, 2010; Harwood & Jones, 2014; Horton & Quesenbery, 2013).

While the majority of attentional studies have been conducted in a laboratory, results from field-based studies suggest that environmental characteristics may play an important role in mobile usability testing (Coursaris & Kim, 2011). In research comparing field and laboratory testing outcomes, researchers identified significantly more usability problems in field tests, which they attributed to the increased cognitive demands of the environment (Duh et al., 2006). Recent work towards the design of intelligent attentional management systems can provide some insight into the importance of field usability testing, as they rely on the collection of large sets of eye tracking and device sensor data (Anderson et al., 2018; Schröder, Hirschl, & Reichl, 2016; Steil et al., 2018). Results from this research show evidence of participants' ability to suppress disruptions by low-level features, as well as the effect of task interruptions. Interruptions are also shown to lead to greater task errors and completion times, where recovery is dependent on the complexity of the distraction and proximity to the original task.

Chapter 3: Methods and data collection

To examine how performance and attention change depending on the type of mobile device used, virtual or physical, a task-based usability study was conducted using a wearable eye tracker in a stationary field environment. Testing sessions were conducted at a local coffee shop with 18 participants recruited on-site (Gender: 9 female, 8 male, 1 nonbinary; Age: m=27.3 SD=7.8). This space provided a moderate level of control for recording, while also introducing some natural background activity. A wearable eye tracker, Tobii Pro Glasses 2, was used to measure the attention given to the device, task, and outside influences. Due to the nature of using this tool, volunteers who were unable to view a laptop or mobile device without the aid of prescription lenses were excluded from this experiment. Research goals focused on the effect of the type of device (virtual or physical) in relation to task performance and success, website strategies, management of off-screen distractions, and perceived usability.

For each session, participants were asked to complete a series of five tasks on two similar websites, using the desktop emulation of a mobile device for one and a physical mobile device for the other. Site selection focused on identifying a type of website which addresses a common activity for mobile device users, online shopping (Smith & Anderson, 2016), but in a context that is less familiar. The recent emergence and growth of online grocery shopping websites (Nielsen & FMI, 2017) provided the opportunity to compare similar mobile-friendly designs, while reducing risk of participants having prior experience with the site. Sessions were conducted over a period of two weeks, during which there were no significant changes to the site design. An iPhone 7 Plus, reset to factory settings, was used as the physical and virtual device in all testing sessions, to ensure comparability of performance between participants. To control for potential carryover effects, the order of stimulus, device, and task sets were also varied across the group using a latin square design (Bojko, 2013).

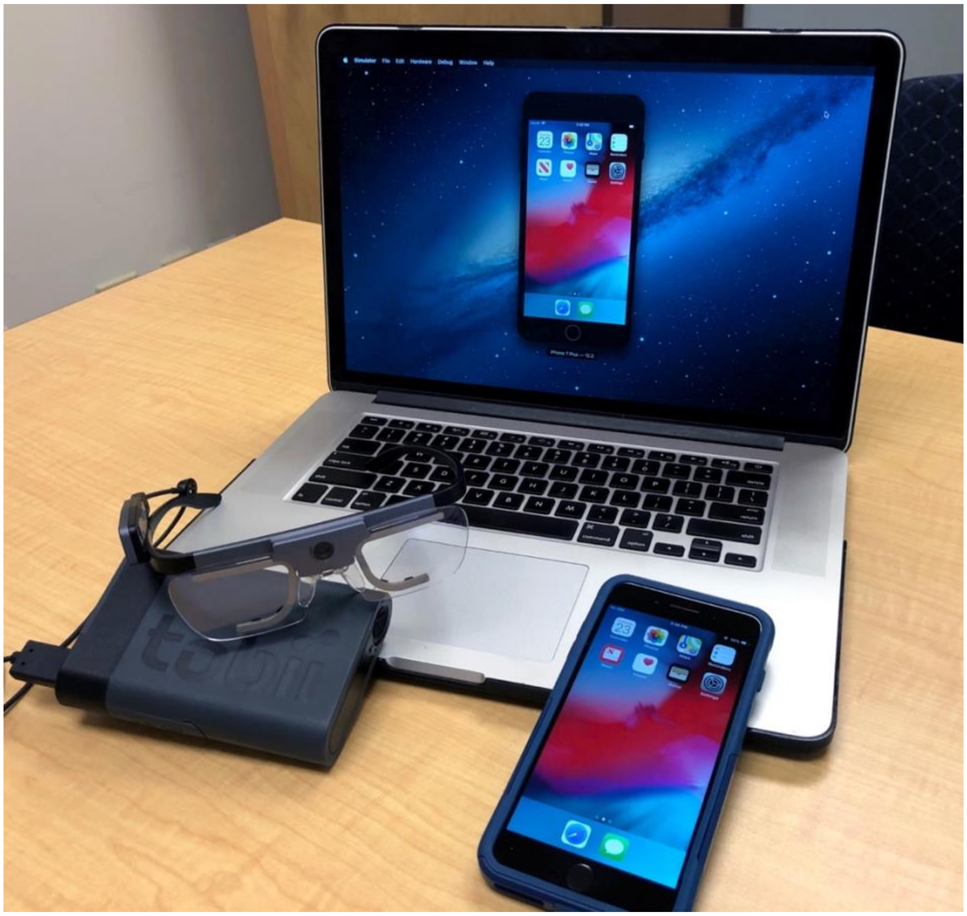

Figure 1. Tobii Pro Glasses 2 eye tracker and testing devices used.

Upon completion of the informed consent, participants were fitted with the Tobii Pro Glasses 2, and the device was calibrated to the participant using the Tobii calibration card and a tablet device running the Tobii Pro Glasses Controller software. The initial testing device was placed in front of the them and each participant was given an opportunity to familiarize themselves with its function. Afterwards, the web browser cache was cleared, the recording was started, and the first task was handed to the participant. Once all five tasks were completed for the first website, participants were asked to completed a standardized usability questionnaire to measure perceived usability of the device and website. The System Usability Scale was selected as it is an established tool for measuring usability and is easy to administer (Sauro & Lewis, 2016). Then the device and website were replaced and the process was repeated with the second website.

Table 1

Overview of task objectives

	Goal
Task 1	Identify an affordable protein / vegetable for a meal.
Task 2	Find all items needed to prepare a recipe and add them to the shopping cart.
Task 3	Replace an item in the recipe to account for a dietary restriction.
Task 4	Select a breakfast item / dessert under a specified price.
Task 5	Change pickup / delivery settings and prepare for checkout.

Note. Two versions of each task were created.

After completion of the testing sessions, recordings for 17 participants were imported into Tobii Pro Lab Analyzer Edition for review. A single participant (P15) was excluded from further analysis, due to a hardware capturing error that occurred during the recording of their session. For each recording, the fixation data was mapped onto a static reference image using the application’s real-world mapping tool. This process allowed for consistent areas of interest (AOIs) to be mapped for quantitative comparison of on-screen and off-screen fixations in each of the participant sessions. To improve the accuracy of this automated image analysis, the reference image was selected for each participant from their longest task while viewing the website home page. To validate the resulting image map, each fixation on the mapped image was visually inspected and compared to the original video for accuracy.

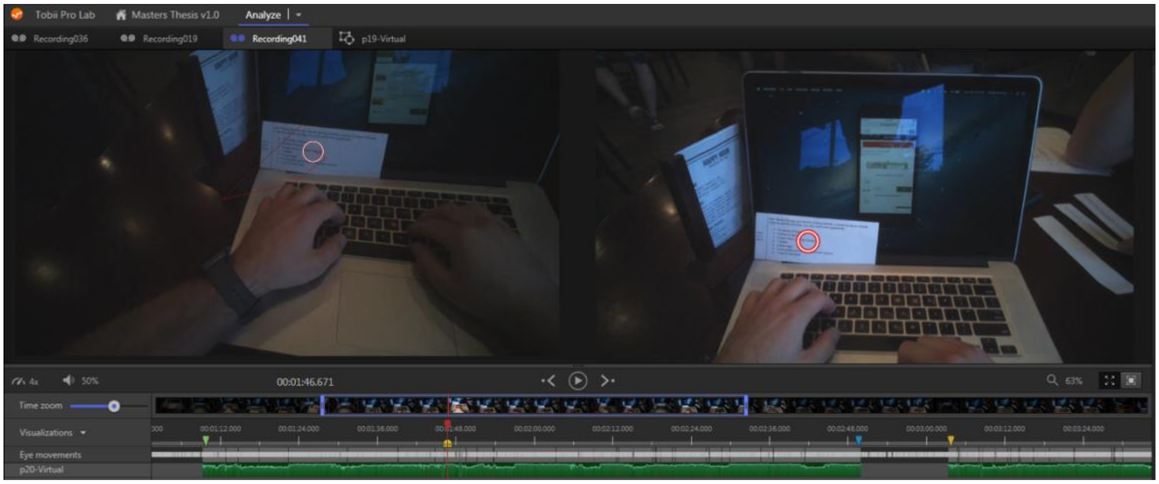

Figure 2. Fixation mapping and review process in Tobii Pro Lab. Each fixation in the original recording (left) was compared to the reference image (right).

Areas of Interest were drawn on each reference image to determine time spent attending to on-screen and off-screen influences, and to allow for a quantitative comparison between the virtual and physical devices. To examine differences in on-screen behavior the total phone screen area was divided into thirds, as changes in viewing behavior between screen resolutions have been identified in prior viewport research (Botella et al., 2014; Fichter & Wisniewski, 2016; Kim et al., 2015; Lagun, Hsieh, Webster, & Navalpakkam, 2014). Additional off-screen areas were calculated for attention given to the instructions and keyboard after the start of a task. To ensure comparability between the use of a physical and virtual keyboard, on-screen fixations on the virtual keyboard were separated from all other physical mobile device data. Metrics were collected for each area of interest including total fixation duration, total fixation count, average visit duration, and visit count. For off-screen areas of interest, the time to first fixation was also collected.

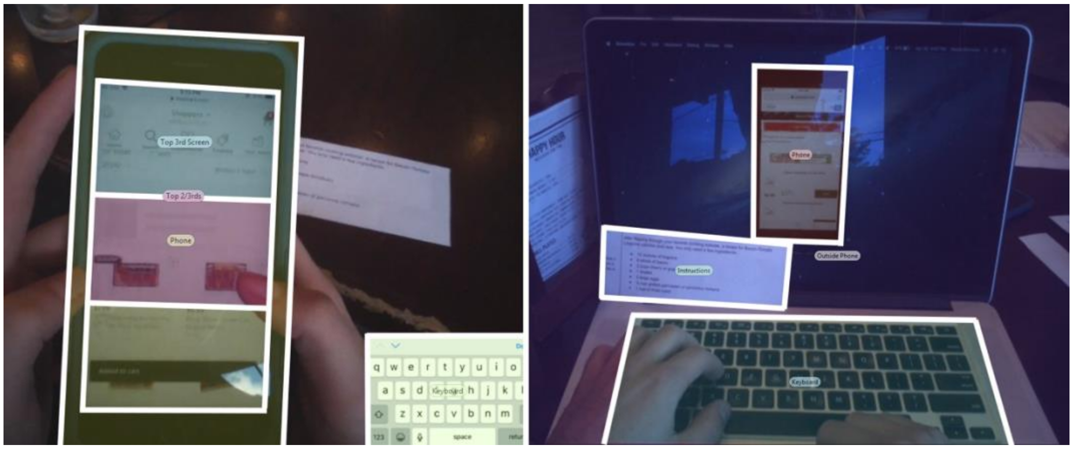

Figure 3. Example of on-screen and off-screen areas of interest.

Table 2

Area of Interest (AOI) measures

AOI	Definition
Total Fixation Duration	The total time each participant has fixated on an area.
Total Fixation Count	The total number of fixations within an area.
Time to First Fixation	The time from the start of the task until an area is first visited.
Average Visit Count	The total number of times an area is visited during a task.
Average Visit Duration	The average duration of a visit to an area.

In addition to examining areas of interest for differences between the virtual and physical device testing, each task was reviewed for time on task, task success, and qualitative observations. Time on task was determined starting with the first fixation on the device after receiving the current task instructions and ending when the task was completed or, in the case of a task failure, when the participant announced their completion. Recordings were also reviewed to identify user strategies, such as browsing product categories or using the search bar, and the use of other website features including sorting, filters, menus, product pages. Events of off-screen attention were also logged for further analysis.

Chapter 4: Results

Performance and satisfaction

Overall, participants were able to successfully complete testing tasks on both websites with minimal issues, regardless of the type of device used. Results show no significant difference in performance between devices as measured by task success, however participants tended to complete tasks faster on a physical device regardless of which device was used first. Similarly, results of the system usability scale survey show a 10.6% improvement in scores when the site was used with a physical device. While the virtual device looked like an exact replica of the physical device, these effects may be attributed to changes in how participants were able to interact with the device. Observations from testing sessions show that several participants encountered difficulties in using a mouse or trackpad to interact with the virtual phone screen, most notably when attempting to scroll through page results. Additionally, some participants attempted to physically interact with the virtual device, suggested that future emulation testing would benefit from the use of a touch screen monitor.

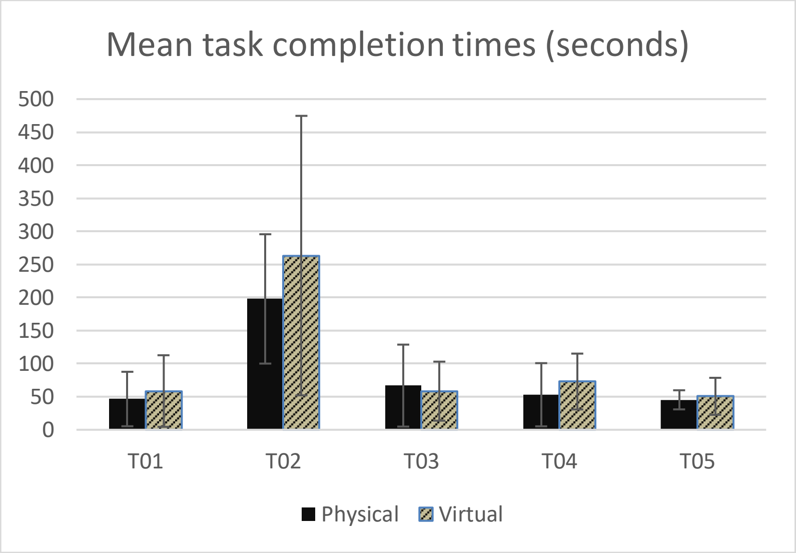

Figure 4. Mean completion times (seconds) by task and device

Table 3

Perception and performance by device

		Physical		*Virtual*		Mean
	n	Mean	SD	Mean	SD	Difference
System Usability Scale	18	74.444	17.459	66.944	17.692	7.5
Time on Task (seconds)	17	409.399	184.365	503.746	263.360	-94.348
Task Success	17	4.588	0.618	4.647	0.702	-0.059

Note. Results in this table not considered significant.

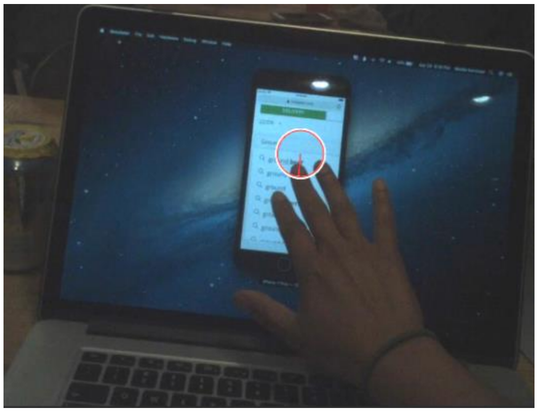

Figure 5. Participant attempts to use the virtual device screen as a touch screen.

Similar to overall task performance, there was minimal change in strategies used by participants to complete tasks on each device. Participants used several features of the website to complete tasks including browsing product categories, using sorting and filters, viewing product pages, and managing the shopping cart, however the search bar was the dominant approach. In addition, participants also opted for search tools whenever a problem was encountered using a browsing approach. This behavior is similar to observations recorded in other mobile usability research (Groth & Haslwanter, 2016). There was also no apparent difference in search depth between devices, however a difference in the layout and design of search results appears to have influenced the decision to visit product pages on one of the websites. Site A's search results feature a large “Add to Cart” buttons whereas Site B used a smaller plus icon to signify the same function. When navigating Site B, many participants opted to visit product pages which also included a larger button that was similar to Site A’s search results.

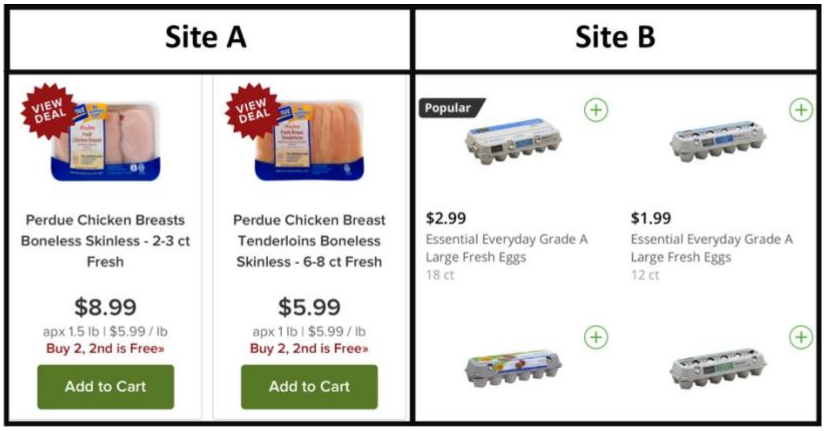
Site A
Site B
VIEW DEAL
Perdue Chicken Breasts Boneless Skinless - 2-3 ct Fresh
$8.99
apx 1.5 lb | $5.99 / lb
Buy 2, 2nd is Free»
Add to Cart
VIEW DEAL
Perdue Chicken Breast Tenderloins Boneless Skinless - 6-8 ct Fresh
$5.99
apx 1 lb | $5.99 / lb
Buy 2, 2nd is Free»
Add to Cart
Popular
$2.99
Essential Everyday Grade A Large Fresh Eggs
18 ct
$1.99
Essential Everyday Grade A Large Fresh Eggs
12 ct

Figure 6. Difference in search result pages.

Table 4

Task Strategy by device

		Physical		*Virtual*		Mean
	n	Mean	SD	Mean	SD	Difference
First Strategy:						
• Search	17	3.059	0.899	3.250	0.754	-0.191
• Browse	17	0.765	1.147	0.529	0.717	0.235
Features Used:						
• Sort / Filter	17	0.353	0.606	0.412	0.507	-0.059
• Product Page	17	1.353	0.966	1.421	1.004	-0.059
• Menu / Cart	17	1.176	1.344	1.529	1.125	-0.353

Note. First strategy denotes the number of tasks where the participant began with the listed approach. Features used denotes the number of tasks in which the website feature was used.

Eye tracking

Analysis of eye tracking measures showed no significant difference in the number of fixations and frequency of visit to all on-screen areas of interest. There were, however, noted differences in the amount of time that attention was given to specific areas on-screen and off-screen, suggesting that type of device used in testing can influence attention. On average, participants spent 49.3% more time attending to the virtual keyboard on the physical device when compared with using the laptop keyboard on the virtual device. The top third of the viewport also received 23.1% more attention when using the virtual device, and average visits to this area were 9.8% longer. Additionally, even though many participants held the physical device in closer proximity to themselves than the virtual device, the average first instance of off-screen attention was 25.4% earlier on physical devices, and the average visit duration for off-screen influences was 33.8% longer.

Table 4

Total Fixation Duration (seconds) by device

		Physical		*Virtual*		Mean
	n	Mean	SD	Mean	SD	Difference
Area Outside Phone	17	33.081	15.442	34.699	25.528	-1.618
Instructions	17	27.491	14.491	24.319	19.232	3.172
Keyboard	17	31.372	25.268	18.959	14.469	12.412*
Phone	17	298.293	146.247	384.177	219.551	-85.884
Top 1/3 Screen	17	213.665	105.886	269.502	154.199	-55.837*
Top 2/3 Screen	17	72.607	36.222	110.887	75.587	-38.280

Note. Virtual and Physical keyboard data has been excluded from Area Outside Phone and Phone. p < .05*

Table 5

Average Visit Duration (seconds) by device

		Physical		*Virtual*		Mean
	n	Mean	SD	Mean	SD	Difference
Area Outside Phone	17	6.514	3.401	4.630	2.165	1.884*
Instructions	17	6.240	3.384	4.717	2.734	1.523
Keyboard	17	4.698	2.510	3.616	2.225	1.082
Phone	17	43.744	25.298	50.327	34.900	-6.583
Top 1/3 Screen	17	11.083	3.906	15.233	9.238	-4.151*
Top 2/3 Screen	17	5.107	1.216	7.104	2.127	-1.997

*Note. Virtual and Physical keyboard data has been excluded from Area Outside Phone and Phone. * p < .05*

In addition to analysis of areas of interest for all task durations, each recording was reviewed in its entirety for off-screen attention during and after each task. A total of 45 events were logged from all recording sessions, with the majority of events occurring during or after Task 2 (20 events), suggesting that the length and repetitive nature of this task (searching for items in a recipe) may have influenced participant’s ability to maintain attention. The most common sources of disruptions included the motion and sounds of employees and customers in café, however several participants also paused to consume a nearby beverage. In most instances, participants would attend to the source of the disruption before returning to the screen or task instructions. These events often occurred between the search and selection of individual items in the task and were often brief. However, a more exaggerated effect was observed post-task which, in one instance, led the participant to repeat steps in the task that had already been completed.

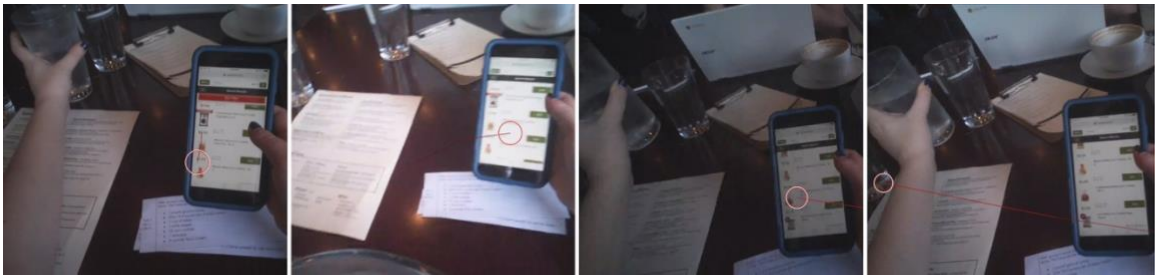

Figure 7. While working on task 2, participant reaches for a drink of water.

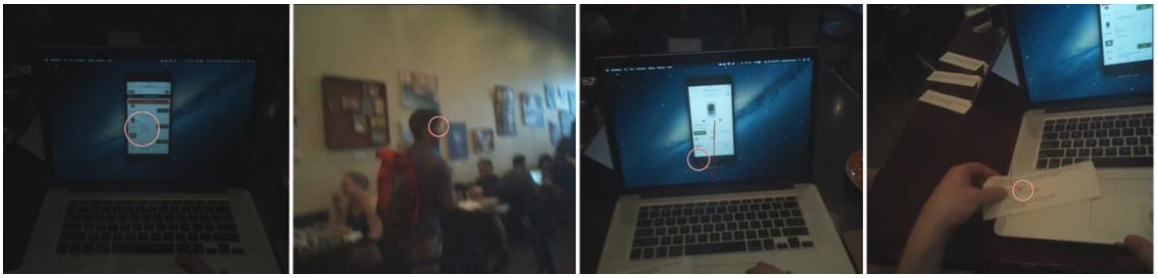

Figure 8. After completing task 2, participant noticed a person walking by then returned to complete the final step in the task again.

Observations of off-screen attention in other tasks may provide insights into participant's thinking and problem-solving strategies. Within task 3, for example, a number of participants exhibited an initial gaze pattern which aligns to established thinking-related gaze behaviors (McCarthy, Lee, Itakura, & Muir, 2008). Task 3 required participants to select a replacement ingredient for the recipe in Task 2, which would address a specified dietary restriction. After starting this task, several participants were observed looking off-screen (up and to right) prior to determining a task strategy, and one participant confirmed this effect by talking through their thought process while performing this gaze pattern. Similar patterns were also observed in other tasks when participants were confronted with an unexpected result, prior to restructuring a search query.

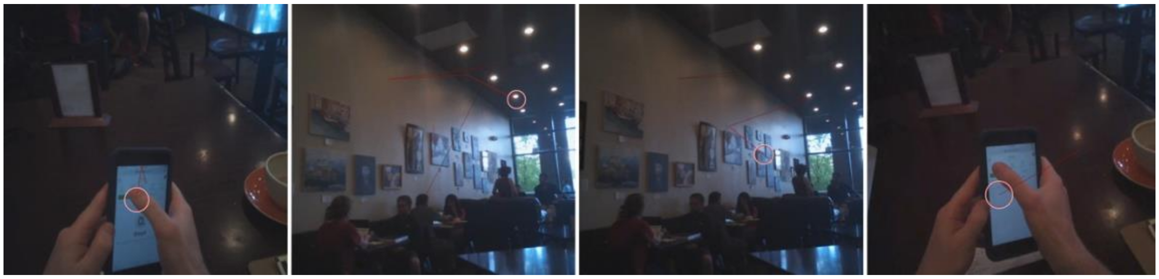

Figure 9. Participant looks off screen while thinking about a search query for Task 3.

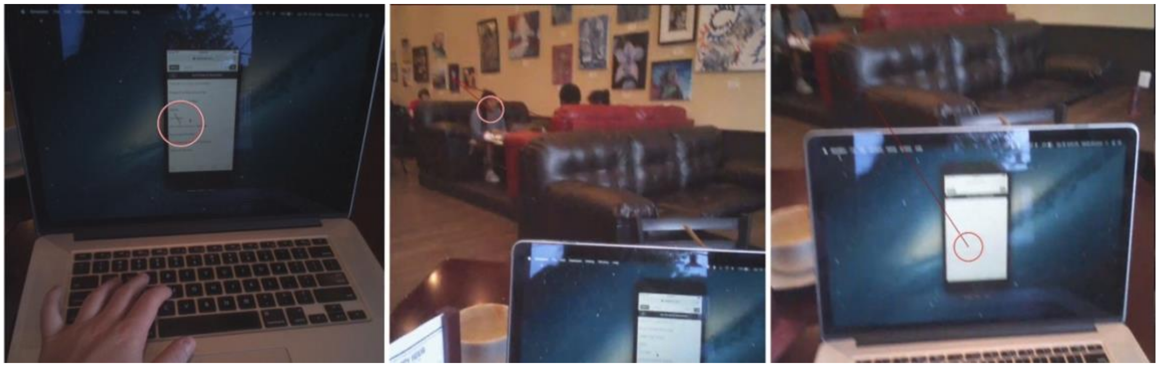

Figure 10. Participant looks offscreen in task 4 when search results do not display a list of products as expected.

Chapter 5: Discussion, limitations, and conclusion

Discussion

This study explores the relationship between the type of device used in relation to mobile usability testing outcomes in a field environment. Results suggest that when compared to a virtual device (emulator), the use of a physical device has an influence on improving time on task and perceived usability. While the use of emulators allow participants to experience a simulation of the physical mobile device, limitations in interface controls (use of mouse or trackpad) were also observed to cause frustration and delays in task completion. Participants also attempted to interact with the virtual device as they would with a physical phone, suggesting that in the future emulation testing should include touch screen capable devices in order to better mimic a mobile device.

The use of eye tracking in this study uncovered aspects of user engagement and attention which would not have been observed using standard performance metrics alone. As the size of the interface viewport was controlled, there were no significant differences in the total number of fixations, however there were measured effects on the duration and location of fixations based on the type of device. Analysis of eye tracking measures identify a bias to the top third of the viewport on a virtual device (emulator) when compared to testing with a physical device. Use of a physical device was also confirmed to significantly increase the time and attention spent to the virtual keyboard on the screen, as well as with off-screen influences. These results expand upon previously observed differences in search patterns and behaviors between physical and virtual devices that have been conducted in a laboratory setting.

As with prior attentional research, the use of task-based testing appears to direct attentional capacity to features (screen, keyboard, and instructions) that were relevant to the current objective. Through analysis of the session recordings, participants were shown to be more susceptible to low-level salient features in-between tasks and steps, and when participants were confronted with an unexpected outcome. When distracted, task recovery was also observed to be dependent on the complexity of the distraction and proximity to the original task. Repetitive and challenging tasks were also shown to increase the chance of distractions, leading to increased task errors and longer completion times.

While there were measurable differences in the observed results between the two devices, it is still unclear what effect the type of eye tracker used would have on field usability testing. The testing environment selected for this study could have allowed use of a device-mounted tracker for the virtual device testing, as participants were in a seated position, and may have allowed for a device mounted tracker for actual devices. However, device-mounted eye tracking for handheld devices has not proven very reliable, nor does it allow for natural handling of the device by the participant. It was important to use the same eye tracking method for both the physical and virtual devices, in order to compare the data meaningfully.

The wearable eye tracker was well suited for comparing the use of emulators and physical devices in a field environment, however quantitative analysis of wearable eye tracking data continues to be a significantly more manual and time-consuming process when compared to device-mounted tracking. Future studies should consider comparing the effect of the type of eye tracking device used in relation to usability outcomes, as usability testing must often weigh the time, costs, and benefits of the testing approach. For example, a study comparing the data collected by a device-mounted tracker to the data collected by a wearable eye tracker would be valuable.

Limitations

While the results of the system usability scale suggest a preference to the physical device, and mean task completion times also suggest improved performance, statistical analysis of these metrics was not found to be significant at the current sample size. This is a common challenge in many usability eye tracking studies, as conducting and analyzing session data can be labor intensive and time consuming. In particular for this study, the testing window of two weeks and the availability of research resources did not allow for a larger sample to be collected. As these metrics do not directly rely on eye tracking measures, a larger follow-up study could be conducted without the use of the wearable eye tracker to further investigate these trends.

Website selection and task design were also constrained by the available testing window for this study. To allow for comparison of both stimuli in a single session, two grocery shopping websites were selected and two similar task sets were designed. While order of exposure was varied across the participants, the length of the session and similarities between website and task sets may have influenced task performance in the second half of each session. Additionally, subtle differences in the designs of the two sites could have influenced performance outcomes. Alternatively, a future study could consider comparing a single cached website and single task set in two independent sessions in order to ensure strong comparability between the emulator and physical device conditions.

Conclusion

This study confirms that wearable eye tracking can be used to collect valuable quantitative and qualitative data in field usability testing. Building on prior research conducted in a laboratory setting, analysis of areas of interest show evidence of changes in attention when comparing the use of a virtual (emulator) or physical mobile device in a usability testing scenario, even as available screen space remains constant between these conditions. Collection of on-screen and off-screen gaze data also provides insight into users' attentional management and thinking strategies, which can help researchers in identifying and correcting usability issues with their mobile designs. Future studies should continue to examine the use of device-mounted and wearable eye trackers with mobile devices in a variety of field settings. As hardware and image analysis continue to improve, the use of wearable eye tracking could also expand into other user research applications including paper prototype testing, card sorting, and contextual inquiry.

References

Almeida, S., Veloso, A., Roque, L., & Mealha, Ó. (2011). The Eyes and Games: A Survey of Visual Attention and Eye Tracking Input in Video Games. https://doi.org/10.13140/rg.2.1.2341.3527

Anderson, C., Hübener, I., Seipp, A.-K., Ohly, S., David, K., & Pejovic, V. (2018). A Survey of Attention Management Systems in Ubiquitous Computing Environments. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2(2), 1–27. https://doi.org/10.1145/3214261

Benjamins, J. S., Hessels, R. S., & Hooge, I. T. C. (2018). Gazecode: open-source software for manual mapping of mobile eye-tracking data. Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications - ETRA '18, 1–4. https://doi.org/10.1145/3204493.3204568

Biedert, R., Dengel, A., Buscher, G., & Vartan, A. (2012). Reading and estimating gaze on smart phones. Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12, 385. https://doi.org/10.1145/2168556.2168643

Bojko, A. (2013). Eye tracking the user experience: a practical guide to research. Brooklyn, New York: Rosenfeld Media.

Borys, M., Czwórnóg, M., & Ratajczyk, T. (2016). Web analytics combined with eye tracking for successful user experience design: A case study. Applied Computer Science, 12(4), 96–110.

Borys, M., & Milosz, M. (2015). Mobile application usability testing in quasi-real conditions. 2015 8th International Conference on Human System Interaction

(HSI), 381–387. https://doi.org/10.1109/HSI.2015.7170698

Botella, F., Moreno, J. P., & Peñalver, A. (2014). How efficient can be a user with a tablet versus a smartphone? Proceedings of the XV International Conference on Human Computer Interaction - Interacción '14, 1–9. https://doi.org/10.1145/2662253.2662317

Cheng, S. (2011). The research framework of eye-tracking based mobile device usability evaluation. Proceedings of the 1st International Workshop on Pervasive Eye Tracking & Mobile Eye-Based Interaction - PETMEI '11, 21. https://doi.org/10.1145/2029956.2029964

Coursaris, C. K., & Kim, D. J. (2011). A Meta-Analytical Review of Empirical Mobile Usability Studies. 6(3), 55.

Cuadrat Seix, C., Veloso, M. S., & Soler, J. J. R. (2012). Towards the validation of a method for quantitative mobile usability testing based on desktop eyetracking. Proceedings of the 13th International Conference on Interacción Persona-Ordenador - INTERACCION '12, 1–8. https://doi.org/10.1145/2379636.2379684

Duh, H. B.-L., Tan, G. C. B., & Chen, V. H. (2006). Usability Evaluation for Mobile Device: A Comparison of Laboratory and Field Tests. Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services – MobileHCI '06, 181–186. https://doi.org/10.1145/1152215.1152254

Dumais, S. T., Buscher, G., & Cutrell, E. (2010). Individual differences in gaze patterns for web search. Proceeding of the Third Symposium on Information Interaction in

Context - IIiX ’10, 185. https://doi.org/10.1145/1840784.1840812

Fichter, D., & Wisniewski, J. (2016). Beyond Responsive Design. Online Searcher, 40(6), 66–68.

Foulsham, T., & Kingstone, A. (2012). Goal-driven and bottom-up gaze in an active real-world search task. Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA ’12, 189. https://doi.org/10.1145/2168556.2168590

Groth, A., & Haslwanter, D. (2016). Efficiency, effectiveness, and satisfaction of responsive mobile tourism websites: a mobile usability study. Information Technology & Tourism, 16(2), 201–228. https://doi.org/10.1007/s40558-015-0041-0

Harwood, T., & Jones, M. (2014). Mobile Eye-Tracking in Retail Research. In M. Horsley, M. Eliot, B. A. Knight, & R. Reilly (Eds.), Current Trends in Eye Tracking Research (pp. 183–199). https://doi.org/10.1007/978-3-319-02868-2_14

Holmberg, N., Sandberg, H., & Holmqvist, K. (2014). Advert saliency distracts children’s visual attention during task-oriented internet use. Frontiers in Psychology, 5. https://doi.org/10.3389/fpsyg.2014.00051

Horton, S., & Quesenbery, W. (2013). A web for everyone: designing accessible user experiences. Brooklyn, New York: Rosenfeld Media.

Johnson, J. (2014). Designing with the mind in mind simple: simple guide to understanding user interface design guidelines (Second edition). Amsterdam ; Boston: Elsevier, Morgan Kaufmann is an imprint of Elsevier.

Khamis, M., Eiband, M., Zürn, M., & Hussmann, H. (2018). EyeSpot: Leveraging Gaze to Protect Private Text Content on Mobile Devices from Shoulder Surfing. Multimodal Technologies and Interaction, 2(3), 45. https://doi.org/10.3390/mti2030045

Kim, J., Thomas, P., Sankaranarayana, R., Gedeon, T., & Yoon, H.-J. (2015). Eye-tracking analysis of user behavior and performance in web search on large and small screens: Eye-Tracking Analysis of User Behavior and Performance in Web Search on Large and Small Screens. Journal of the Association for Information Science and Technology, 66(3), 526–544. https://doi.org/10.1002/asi.23187

Kuhnel, M., Seiler, L., Honal, A., & Ifenthaler, D. (2017). Mobile Learning Analytics in Higher Education: Usability Testing and Evaluation of an APP Prototype. International Association For Development Of The Information Society, 8.

Lagun, D., Hsieh, C.-H., Webster, D., & Navalpakkam, V. (2014). Towards better measurement of attention and satisfaction in mobile search. Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR '14, 113–122. https://doi.org/10.1145/2600428.2609631

Lazar, J., Feng, J. H., & Hochheiser, H. (2010). Research methods in human-computer interaction. Chichester, West Sussex, U.K: Wiley.

Levulis, S. J., & Harris, D. J. (2015). Are All Tests Equal? A Comparison of Emulator and Device Testing for Mobile Usability Evaluation. Proceedings of the Human

Factors and Ergonomics Society Annual Meeting, 59(1), 976–980. https://doi.org/10.1177/1541931215591281

Magnussen, R., Zachariassen, M., Kharlamov, N., & Larsen, B. (2017). Mobile Eye Tracking Methodology in Informal E-Learning in Social Groups in Technology-Enhanced Science Centres. 15(1), 13.

McCarthy, A., Lee, K., Itakura, S., & Muir, D. W. (2008). Gaze Display When Thinking Depends on Culture and Context. *Journal of Cross-Cultural Psychology*, *39*(6), 716–729. https://doi.org/10.1177/0022022108323807

Meißner, M., Pfeiffer, J., Pfeiffer, T., & Oppewal, H. (2017). Combining virtual reality and mobile eye tracking to provide a naturalistic experimental environment for shopper research. Journal of Business Research. https://doi.org/10.1016/j.jbusres.2017.09.028

Meschtscherjakov, A., Wilfinger, D., & Tscheligi, M. (2014). Mobile attachment causes and consequences for emotional bonding with mobile phones. Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14, 2317–2326. https://doi.org/10.1145/2556288.2557295

Nielsen, & FMI. (2017). The digitally engaged food shopper. Retrieved from https://www.fmi.org/forms/store/ProductFormPublic/the-digitally-engaged-food-shopper

Nogueira, T. C., & Ferreira, D. J. (2016). Evaluating the Impact of Responsive and Non-responsive Web Design on the Experience of Blind Users. IEEE MULTIMEDIA, 8.

Ohm, C., Müller, M., & Ludwig, B. (2017). Evaluating indoor pedestrian navigation interfaces using mobile eye tracking. Spatial Cognition & Computation, 17(1–2), 89–120. https://doi.org/10.1080/13875868.2016.1219913

Pearson, R., & van Schaik, P. (2003). The effect of spatial layout of and link colour in web pages on performance in a visual search task and an interactive search task. International Journal of Human-Computer Studies, 59(3), 327–353. https://doi.org/10.1016/S1071-5819(03)00045-4

Sauro, J., & Lewis, J. R. (2016). Quantifying the user experience: practical statistics for user research (2nd edition). Cambridge: Morgan Kaufmann.

Schröder, S., Hirschl, J., & Reichl, P. (2016). CoConUT: context collection for non-stationary user testing. Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct - MobileHCI '16, 924–929. https://doi.org/10.1145/2957265.2962658

Smith, A., & Anderson, M. (2016, December 19). Online Shopping and E-Commerce | Pew Research Center. Retrieved April 10, 2019, from https://www.pewinternet.org/2016/12/19/online-shopping-and-e-commerce/

Steil, J., Müller, P., Sugano, Y., & Bulling, A. (2018). Forecasting user attention during everyday mobile interactions using device-integrated and wearable sensors. Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '18, 1–13.

https://doi.org/10.1145/3229434.3229439

Tamber-Rosenau, B. J., & Marois, R. (2016). Central attention is serial, but midlevel and peripheral attention are parallel—A hypothesis. Attention, Perception, & Psychophysics, 78(7), 1874–1888. https://doi.org/10.3758/s13414-016-1171-y

Tupikovskaja-Omovie, Z., & Tyler, D. (2018). Mobile consumer shopping journey in fashion retail: eye tracking mobile apps and websites. Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications - ETRA '18, 1–3. https://doi.org/10.1145/3204493.3208335

Appendix

Table 6

Participant and stimulus order

	Age	Gender	First Stimulus	Second Stimulus
P01	24	Female	Virtual (Site A, Task Set 1)	Physical (Site B, Task Set 2)
P02	25	Female	Virtual (Site B, Task Set 2)	Physical (Site A, Task Set 1)
P03	26	Male	Physical (Site B, Task Set 2)	Virtual (Site A, Task Set 1)
P04	30	Male	Physical (Site A, Task Set 1)	Virtual (Site B, Task Set 2)
P05	24	Female	Virtual (Site A, Task Set 2)	Physical (Site B, Task Set 1)
P06	24	Nonbinary	Virtual (Site B, Task Set 1)	Physical (Site A, Task Set 2)
P07	24	Male	Physical (Site A, Task Set 2)	Virtual (Site B, Task Set 1)
P08	26	Male	Physical (Site B, Task Set 1)	Virtual (Site A, Task Set 2)
P09	41	Male	Virtual (Site A, Task Set 1)	Physical (Site B, Task Set 2)
P10	40	Female	Virtual (Site B, Task Set 2)	Physical (Site A, Task Set 1)
P11	21	Female	Physical (Site B, Task Set 2)	Virtual (Site A, Task Set 1)
P12	19	Female	Physical (Site A, Task Set 1)	Virtual (Site B, Task Set 2)
P13	36	Female	Virtual (Site A, Task Set 2)	Physical (Site B, Task Set 1)
P14	42	Female	Virtual (Site B, Task Set 1)	Physical (Site A, Task Set 2)
P15	22	Male	Physical (Site A, Task Set 2)	Virtual (Site B, Task Set 1)
P16	21	Female	Physical (Site B, Task Set 1)	Virtual (Site A, Task Set 2)
P19	21	Female	Physical (Site A, Task Set 2)	Virtual (Site B, Task Set 1)
P20	23	Male	Physical (Site B, Task Set 1)	Virtual (Site A, Task Set 2)

Table 7

Fixation Count

		Physical		*Virtual*		Mean
	n	Mean	SD	Mean	SD	Difference
Area Outside Phone	17	80.059	40.485	81.941	69.391	-1.882
Instructions	17	60.647	20.304	51.471	43.747	9.176
Keyboard	17	59.588	32.137	68.647	44.359	-9.059
Phone	17	540.412	216.440	555.235	368.571	-14.824
Top 1/3 Screen	17	368.118	147.151	376.647	262.816	-8.529
Top 2/3 Screen	17	128.471	58.430	161.882	135.011	-33.412

Note. Virtual and Physical keyboard data has been excluded from Area Outside Phone and Phone. Results in this table not significant.

Table 8

Visit count

		Physical		*Virtual*		Mean
	n	Mean	SD	Mean	SD	Difference
Area Outside Phone	17	33.353	21.166	41.412	35.260	-8.059
Instructions	17	23.235	8.288	20.882	16.800	2.353
Keyboard	17	28.000	13.910	25.765	9.004	2.235
Phone	17	59.000	23.736	59.647	33.621	-0.647
Top 1/3 Screen	17	115.765	50.700	115.941	63.039	-0.176
Top 2/3 Screen	17	71.941	26.267	84.294	48.391	-12.353

Note. Virtual and Physical keyboard data has been excluded from Area Outside Phone and Phone. Results in this table not significant.

Table 9

Time to first fixation (seconds)

		Physical		*Virtual*		Mean
	n	Mean	SD	Mean	SD	Difference
Area Outside Phone	17	49.285	43.895	63.643	65.998	-14.359
Instructions	17	56.061	47.094	55.991	40.829	0.070
Keyboard	17	77.234	89.562	76.644	129.374	0.590

Note. Virtual and Physical keyboard data has been excluded from Area Outside Phone and Phone. Results in this table not significant.