Handheld Augmented Reality: Does Size Matter?

Lawrence Sambrooks and Brett Wilkinson
School of Computer Science, Engineering and Mathematics
Flinders University
PO Box 2100, Adelaide 5001, South Australia
{lawrence.sambrooks,brett.wilkinson}@flinders.edu.au

Abstract

Handheld devices have become extremely popular in recent years and represent attractive options for augmented reality (AR) research. Most modern devices now incorporate many of the necessary input and output capabilities and do so in self-contained packages of varying size, weight, and cost. But while most previous AR work with handhelds has focused on smaller form factors, we have been interested in further exploring the range of larger devices often referred to under the umbrella of 'tablets'.

This paper presents the results from a study we conducted on the suitability of different form factors for mobile AR use. Three form factor categories were evaluated: smartphone, mini tablet, and tablet. Although most devices today are marketed as being either the first or last, we propose there needs to be a third, middle category that caters for the subtle differences between sizes.

The study asked 15 participants to use a device from each category to complete a series of seven interactive tasks. The tasks were designed to incorporate typical AR interactions. Participants completed pre- and post-test questionnaires and were audio recorded during the testing process. Our results showed that no one form factor was best suited to all tasks but rather the 'right' form factor was influenced by task specifics and personal preferences. In terms of usability ratings, we found a significant difference between smartphone and tablet form factors but no such difference between other combinations. Finally, we noted a negative correlation between participants' fatigue rating and the ease with which they found completing the tasks.

Keywords: handheld augmented reality; mobile computing; usability

1 Introduction

Augmented reality (AR) is a process in which virtual objects are superimposed onto the real world in real time (Azuma 1997). Modern handheld devices provide an attractive option for delivering 'magic lens' AR experiences as they offer a truly mobile, self-contained form factor incorporating the types of sensors useful for implementing tracking and registration functionality. Pick up any modern mid to high end handheld and you'll likely find its hardware sporting both a front and rear-facing camera, GPS, inertial measurement unit (IMU), Wi-Fi, and even a barometer. And it's not just sensors, CPU and GPU performance has also increased by orders of magnitude. Early handhelds were very much a compromise, sacrificing CPU and GPU power for mobility and some semblance of battery life, whereas recent devices now incorporate CPUs and GPUs with multiple cores and billions of transistors. The rapid evolution of handhelds has opened up plenty of scope for researchers and developers to explore new and previously unattainable ways of providing mobile AR experiences. Dedicated mobile application frameworks such as Wikitude, Layar, and BlippAR enable simple AR applications to be rapidly deployed to the most common mobile platforms. If AR is to become a mainstream medium, handheld devices will almost certainly play a part.

While a large body of work on mobile AR exists around handheld devices that we would classify as fitting the smartphone form factor, there seems to be limited work utilising larger tablet devices. Granted, some researchers treat tablet devices as oversized smartphones believing they have limited applicability for AR given their size and weight (Arth and Schmalstieg 2011). Nevertheless, we are interested in further exploring the use of larger tablet devices and also so-called in-between 'phablet' devices—a portmanteau of the words *phone* and *tablet*—which we refer to as mini tablets. We intend to investigate the effects these different device sizes have on users' engagement with AR content as part of a larger body of work we are undertaking on mobile AR authoring. While there is an obvious trade-off between the physical size and weight of a device versus its screen size, we wish to better understand why one particular form factor might be preferred over another, if there is indeed a 'goldilocks' size, and whether certain form factors are best suited to particular task types.

We devised an experiment to compare AR interaction for three different handheld devices. The devices used were chosen as prototypical representations of three categories: smartphone, mini tablet, and tablet. The classification of these categories and how we chose to differentiate them is discussed in section 3. The experiment itself consisted of a series of interactive tasks that were designed to explore the types of interactions that might often be used in a mobile AR application. Many of these interactions leverage the unique modality capabilities of modern handheld devices, incorporating both touch and sensor-driven input.

This paper contributes to the broader understanding of mobile AR usability and serves as a base of reference for our work on mobile AR authoring. In the sections that follow, we outline the details of our experiment and present our results along with corresponding discussion.

2 Related Work

While the number of comparative studies between recent handheld AR devices is limited, a few researchers have undertaken work to compare devices of different sizes and in different configurations. Dey, Jarvis *et al*. (2012) investigated the differences in egocentric, exocentric, and ordinal depth perception between smartphone and tablet devices. Their experiments involved distance estimations between virtual objects and the participant (egocentric), the closer of two virtual objects (ordinal), and the distance between two virtual objects (exocentric). They found that bigger displays, such as those on tablet devices, do not improve egocentric or exocentric depth perception but do significantly improve ordinal depth perception. For egocentric depth perception in particular, smartphones were discovered to cause less depth compression, which resulted in less underestimation. Anecdotally, participants were noted as subjectively preferring larger screen tablet devices for egocentric depth perception.

Focussing more on interaction, Wither, DiVerdi *et al*. (2007) compared three different display types for AR-based selection and annotation tasks, two of which were handheld displays and one a head-mounted display (HMD). The distinction between the handheld displays was based on whether the device was held at waist level with the camera pointing parallel to the screen or in a 'magic lens' configuration with the camera pointing perpendicular to the screen. They found handheld displays in a 'magic lens' configuration faster for selection tasks than the other two displays. For annotation tasks, there were no significant differences between the displays; however, participants subjectively preferred the handheld displays when the objects being annotated were real and the HMD when they were virtual.

Other researchers have further investigated AR interaction techniques in an attempt to address issues that arise when interaction occurs on the same device that is tracking the environment. Güven, Feiner *et al*. (2006), Lee, Yang *et al*. (2009), and Langlotz, Mooslechner *et al*. (2012) have all implemented approaches whereby the current AR view is 'frozen' by pausing camera input and tracking. While frozen, the device is able to be repositioned to a more comfortable orientation and virtual objects on the screen may be interacted with free from the effects of instability. Once the necessary interactions have occurred, the view may be unfrozen and camera input and tracking resumed. Güven, Feiner *et al*. further presented 'freeze-n-move' and 'freeze-n-link' modes that assisted with repositioning and linking virtual objects within an organisational hierarchy.

As well as offering a similar 'freeze' method, Bai, Lee *et al*. (2012) proposed a 'finger gesture' technique that involved interpreting intangible gestures performed in front of the device's rear-facing camera. More recently, Vincent, Nigay *et al*. (2013) suggested 'shift-and-freeze' and 'relative pointing' techniques. Their 'shift-and-freeze' technique similarly freezes the camera frame while 'relative pointing' stabilises a selection cursor in the object's frame of reference without freezing. While our study does not incorporate any such techniques for task interaction, our results should serve to highlight similar usability issues present in interactions performed in certain orientations, such as those requiring the device to be held out in front of the user.

Kurkovsky, Koshy *et al*. (2012) mention ergonomics as one of the key usability concerns with handheld AR due to the need to frequently stretch out one's hands and arms while holding a device. Most off-the-shelf handheld devices, including those used for this study, have not been designed with AR use in mind and frequently include cameras in undesirable locations or thin bezels that may cause accidental interactions with the screen. Kruijff and Veas (2007, 2008) also identified these shortcomings and developed a purpose-built handheld AR device called Vesp'R. The design was built around the Ultra Mobile PC platform and was specifically engineered to be ergonomic and support prolonged use via the inclusion of joystick-like handles. The handles were designed to avoid accidental occlusion of the screen that results when users try to find a comfortable grip on devices with thin bezels. The joysticks support two configurations: one where both are mounted either side of the display to resemble a steering wheel and the other where a single joystick is mounted centrally below the display.

Though not strictly related to AR, an interesting article by Hoober (2013) discusses the results of a study undertaken to discover how users hold their smartphone while interacting with it. The study made 1333 public observations of users using their device while going about their daily routines. The observations did not try to identify the particular device nor the application that was being used, and there was no count of the total number of people encountered. The results found 49% of users held their device with one hand, 36% cradled—a term used to describe a case in which a device is held with two hands but only one is used for interaction—and the remaining 15% with two hands. Unfortunately, the data does not show the particular task users were performing when the observations were made, which would have helped explain why a device was being held in a particular way. This limitation is acknowledged by Hoober as a practical barrier with public observation.

3 Handheld Form Factors

Our study revolved around comparing devices from three different categories: smartphone, mini tablet, and tablet. The distinction between these categories was a choice we made in the absence of any official or de facto standard. Terms such as smartphone and tablet are widely used by manufacturers to market their products but they aren't necessarily used consistently. This can make it difficult to discuss and compare devices or generalise results. It is also likely one of the reasons for the emergence of the word *phablet* to describe devices that can legitimately be considered both smartphones *and* tablets.

While Wagner, Pintaric *et al*. (2005) also identified three classes of devices for handheld AR use—cellular phones, PDAs, and Tablet PCs—the device types they referred to are now largely out-of-date with respect to current technology. We have undertaken our own categorisation of current devices so we can clarify the boundaries that determine when a device transitions from being one type to another. The categorisation scheme employed in our experiment is by no means our attempt to formalise

a method for categorising handheld devices but rather an approach we have used to help clarify our results and comparisons.

As a starting point, we surveyed the range of devices offered by well-known manufacturers that were marketed as being either a smartphone or tablet. Throughout our sampling, we did not come across any other terms used. In order to be considered, each device had to run a mobile operating system supporting the creation of apps and have at least a rear-facing camera: the minimum specifications necessary to be considered usable for AR. As modern handhelds are essentially 'all screen', screen size was used as the key metric. We only considered devices offered by each manufacturer that had distinct screen sizes, so if five devices were available with four inch screens, we only counted one of them.

Our sample consisted of 90 distinct devices across 14 different manufacturers. Some manufactures produced tablets only, some smartphones only, and others both. We collated the results into screen size groupings of one inch; i.e. three to four inches, four to five inches, and so forth. If a device had a screen size of exactly four inches, it was placed into the four to five inch category. Figure 1 shows the distribution of screen sizes among the devices sampled.

It is clear from the graph that devices with a screen size less than six inches make up a majority of the sample (59%). These devices are all marketed as smartphones and we consider this the smartphone category as, for most people, they are usable with one hand. Anything above six inches starts to move into the realm of tablets and the use of two hands. As mentioned previously, there is little distinction among manufactures between a tablet device with a seven inch screen and one with a ten inch screen: they are all referred to as tablets. We don't consider these devices 'the same' as there are often significant differences in terms of functionality, ergonomics, weight, and intended use. We have therefore sub-categorised tablets into a mini tablet category and a full tablet category. A summary of our categories is presented in Table 1.

Table 1: Form factor categories

Category name	Abbreviation	Screen size
Smartphone	S	3 to 6 inches
Mini tablet	MT	6 to 9 inches
Tablet	T	Greater than 9 inches

4 System Overview

A total of seven tasks were created for the experiment, each designed to make use of different combinations of interactions. Tasks were designed to function like mini-games, each with their own defined goal and rules. The tasks were implemented via a purpose-built application developed on top of the Qualcomm Vuforia AR library (version 2.8.7) using the Unity development environment (version 4.3). The Vuforia library is available as a plugin for Unity and was chosen as it provides a comprehensive feature-set, is free to use, and supports all modern mobile platforms.

Part of the feature-set provided by Vuforia is support for a number of different tracking approaches ranging from those based on fiducial markers to more complex

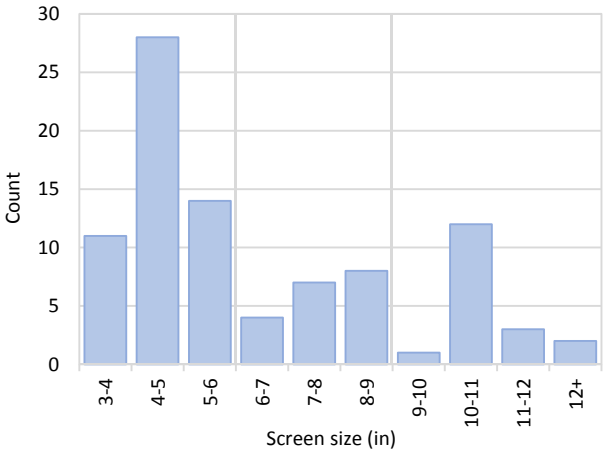

Figure 1: Distribution of screen sizes

computer vision algorithms such as SLAM (Simultaneous Localization and Mapping). We decided to utilise fiducial markers given the devices being used and the anticipated testing environment. The use of fiducial markers for handheld AR is well understood and provides more than acceptable tracking performance. Within the Vuforia library, fiducial markers are referred to as *frame markers*. Instead of utilising a template matching technique to recognise a unique symbol printed on each marker in the way toolkits such as ARToolkit (Kato and Billinghurst 1999) do, frame markers incorporate a unique binary pattern encoded within their borders (Figure 2). The interior area of the marker is left unused and may be filled with custom images or text.

Each task in the experiment is represented by a single marker object with the exception of task seven, which spans two markers. Key to the operation of each task is a custom control script component that is responsible for the task's overall behaviour and interaction. All control scripts inherit from a base script that provides common functionality to all tasks and includes a special loop that we refer to as the *interaction loop*. The interaction loop is invoked whenever a task's marker enters its tracked state. The loop continues to run for as long as the marker remains tracked and the task has not been completed. The purpose of the interaction loop is to control the behaviour of any virtual props used by the task and to process user interactions with it. The loop can be thought of like a miniature game loop that runs independently for each task. This allows each task to be self-contained which in turn allows participants to move between them in any order they choose. We did not wish to enforce on participants a particular way of progressing through the tasks nor did we want each subsequent task to be dependent on the last. Using this approach, it would be trivial to extend the experiment to include additional tasks beyond the seven we created simply by adding a new task marker and associated control script. A top-down illustration of all tasks is presented in Figure 2 (over page) and we discuss each in more detail throughout the remainder of this section.

4.1 Start

To ensure all participants started from the same location, we decided to implement a simple 'tap to start' interaction using a start marker. The start interaction merely required

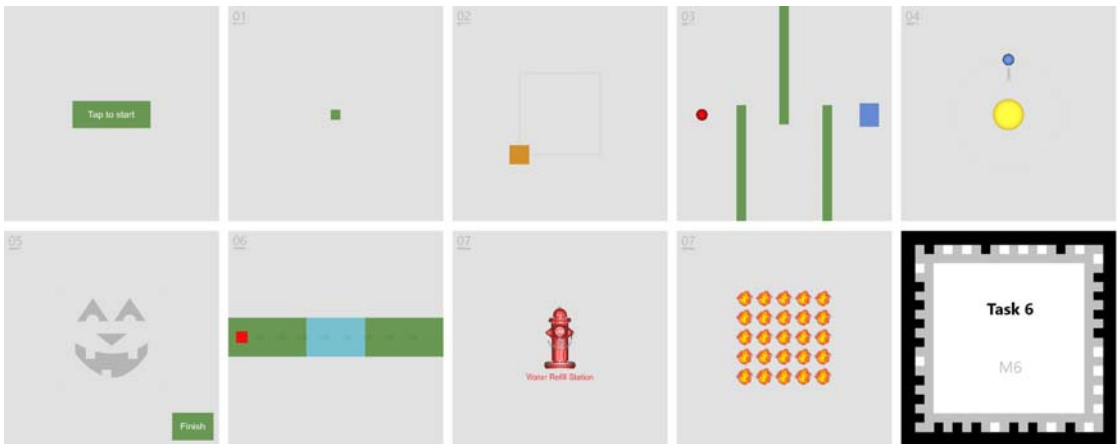

Figure 2: Illustration of tasks and fiducial marker. In order from top left: start task/interaction, Task 1, Task 2, Task 3, Task 4, Task 5, Task 6, Task 7a, Task 7b, Vuforia frame (fiducial) marker

participants tap a virtual button before proceeding to the first task. Architecturally, we made tapping the start button a requirement for the remaining tasks to become trackable (visible) and so this marker/interaction could not be skipped. To assist with our note-taking during the experiment, we also logged the start time for the session and the initial orientation of the device when the button was pressed.

4.2 Task 1

The first task involved no direct screen interaction and was designed to evaluate holding a handheld device steady for a certain period of time. Completing the task required participants to align a crosshair, displayed in the centre of the screen, over a 3D cube (anchored to the marker) and keep it aligned for ten seconds. Any unalignment of the two before this time had elapsed would cause the task to reset. The choice to use ten seconds as the alignment time was arbitrary. However, preliminary testing showed it to be an appropriate choice. By requiring that the device be positioned over the marker for ten seconds, we aimed to determine the effects of device size and weight on the ability to maintain focus as well as observe the way different-sized devices were being held when approaching this type of task.

Informing the user of the time remaining was achieved via a progress bar that appeared along the bottom edge of the screen. The progress bar was designed to allow for peripheral monitoring of progress without distracting the user from the task. To prevent too much variability in the distance participants stood from the marker, and to make the task somewhat less trivial, we also enforced a distance requirement of between 40 and 60 centimetres. The crosshair position would only be registered within this distance range. The marker was placed vertically on a wall at waist level (80 centimetres above the floor) to see how participants would adapt their body and grasp of the device.

4.3 Task 2

The second task was designed to explore how holding a device above eye level would affect screen interactions. We presented participants with an orange button that moved, clockwise, along a square path. To complete the task, participants had to select the button by single tapping on it. Once the button had been successfully selected, it changed colour (to green) and stopped moving. From the participants' point of view, the movement speed of the orange button appeared to be constant; however, we decreased the speed of the button by a fifth of its original speed for every two consecutive misses. The speed reduction was gradual and not intended to be noticeable. We included this feature as a way to subtly assist the participant without any involvement on our part during testing. We did not want to influence participants' thoughts towards a particular device form factor by providing any direct assistance. As we logged all activity, including the number of times this 'feature' was invoked, we do not consider it adversely affected the results of the task but rather helped to highlight usability issues between the form factor being used and the interactions required.

We placed the marker for this task vertically on a wall 177 centimetres above the floor and enforced no distance requirements.

4.4 Task 3

The third task was designed to explore the potential effects of occlusion. Participants were presented with a simple maze and were required to draw a path through the maze starting from a red sphere, representing the player or avatar, to a blue exit area. Drawing the path was accomplished via tracing it on the device's touch screen. As soon as the path was finished, indicated by the drawing finger leaving the screen, the red sphere would begin to follow it at a constant speed. During this time, interaction with the screen was ignored so the path couldn't be altered. Any contact between the red sphere and the walls of the maze were treated as a collision that would cause the task to reset. We were interested to see how participants would approach drawing the path on devices with different sized screens and whether they would alter their grasp of the device or drawing finger to compensate for any occlusion—smaller screens are obviously more prone to occlusion than larger screens and also suffer more from the 'fat finger' effect. We placed the marker for this task horizontally on a table (75 centimetres high) and did not enforce any distance requirements.

4.5 Task 4

The fourth task shared some similarities with the first and second task. The task presented participants with an outer sphere that orbited an inner sphere and was described in

terms of a planet orbiting its sun. Instead of selecting a moving object as in task two, this task required that the device remain focused over the orbiting sphere, similar to task one. We wished to explore the manoeuvrability characteristics of each device form factor and therefore not only orbited the sphere but also gradually reduced its speed while the device remained focussed on it. The current orbit speed was displayed to participants on a label that was anchored to the orbiting sphere and moved with it. Once the orbit speed reached zero (i.e. stopped), the task was considered complete. If the device lost focus while the orbiting sphere was still moving, the orbit speed was rapidly increased back to its original value. We considered a complete task reset too harsh and so this was thought to be a valid compromise. Additionally, if the participant lost focus ten times consecutively, we gradually increased the orbiting sphere's diameter, similar to the approach used in task two.

We placed the marker for this task vertically on a wall at shoulder height (147 centimetres above the floor) and enforced a distance requirement of between 40 and 55 centimetres.

4.6 Task 5

Task five introduced free-form drawing in a non-ideal vertical plane and required participants to trace over the silhouette of a pumpkin face. In designing this task, we were interested in observing how participants would approach drawing in this orientation and also how successful their drawings would be between different devices. Drawing the shapes that made up the face required similar motions to task three where the path of the shape (its outline) would appear on the screen as traced by the participant. Once the outline had been completed, indicated by the tracing finger leaving the screen, the system would take the path data and compute a 2D plane from it, placing it at the exact location where the outline was drawn. Shape outlines were not allowed to intersect and would be ignored if they did. If desired, completed shapes could be repositioned by dragging them with a finger but were not able to be deleted. Once a participant was satisfied with their attempt, they could tap a finish button to complete the task. We placed the marker for this task vertically on a wall at shoulder level (163 centimetres above the floor) and enforced no distance requirements.

4.7 Task 6

While task five was designed to evaluate free-form drawing in a vertical plane, task six explored a more rigid drawing exercise in a horizontal plane, allowing us to compare different orientations. The aim of the task was to draw a bridge between two raised blocks. The bridge could be any shape desired so long as it remained within the boundary extending over and between the two blocks. Drawing was again similar to task three and five. Once the drawing finger left the screen, the bridge outline was converted into a 2D plane. A red cube, used to represent the player or avatar, would then attempt to traverse the gap between the two blocks by following a pre-defined straight-line path at a constant speed. Once the red cube started to move, interaction with the screen was ignored so the bridge couldn't be modified. If the bridge was sufficient, the red cube would successfully cross over it to traverse the gap; if the bridge was insufficient, the red cube would fall and the task would reset. The marker for this task was placed horizontally on a table (75 centimetres high) and there were no distance requirements enforced.

4.8 Task 7

All of the previous tasks were fairly ephemeral in nature, occupying a single location and requiring a short amount of time to complete (barring multiple attempts). For the final task, we wanted to explore an interaction that was more prolonged and not confined to a single location. Task seven therefore required that the participant move between two markers placed approximately 30 metres apart. The task played out like a mini game in which one marker displayed 25 small fires that needed to be extinguished. Extinguishing the fires required water that could be collected from the second marker. Water was represented in terms of *water blocks* of which a maximum of five could be carried at any one time. Participants would navigate to the water marker to collect water and then back to the fires marker to use that water to extinguish the fires; they would then repeat the process until all the fires had been extinguished. While the task may have been a little repetitive, the aim was to see how participants felt using the devices for longer periods of time and whether issues such as fatigue started to become noticeable as they progressed.

While pilot testing the task, we noticed that after collecting water, most users would simply walk back to the fires marker with the device held at their side somewhat nullifying any potential effects of prolonged use and interaction. In response to this, we decided to modify the task to require that the device be held upright, within ±20 degrees, whenever water had been collected in the same way a bucket of water would be carried upright. Tilting the device beyond this threshold would cause water blocks to gradually 'spill', one every two seconds, again mimicking the effect of tilting a bucket of water too much. The current angle of the device was displayed to participants in the form of an icon that appeared in the centre of the screen along with the current device angle. The icon would turn red as soon as the device tilted beyond the threshold and a short beep would sound every time a water block was deducted. If water was spilled, it would need to be replenished from the water marker.

5 User Evaluation

A total of 15 participants were recruited from Flinders University to take part in the study. Of the 15, 14 were male and one was female. The average age range was between 21 and 30. The study made use of a within-subjects design whereby each participant completed the same set of tasks using each device form factor. Upon agreeing to take part, participants were supplied with a copy of a document referred to as the *Tasks Overview Document*. The Tasks Overview Document described each of the tasks in a predominately pictorial manner and explained a suggested method for completing them. The document also outlined important user interface elements that would be required for certain tasks, such as the dis-

tance meter used to monitor distance. As the goal of the experiment was to compare the usability of different device form factors and not the usability of an AR software system, we wanted to ensure participants would be as familiar as possible with the tasks they would be asked to complete before they attempted them. Revealing the tasks beforehand meant the focus could remain on the use of each device rather than on learning the tasks or the software system.

Prior to attempting any tasks, we asked participants to complete a pre-test questionnaire that was designed, in part, to collect background information on handheld devices currently owned. The final three questions of the questionnaire were prefaced with a statement asking participants to re-familiarise themselves with the Tasks Overview Document (a copy was supplied with the questionnaire). These last three questions were designed to ascertain participant preferences with full-knowledge of the devices and tasks. The questions sought a rating of hardware factors that were believed to be important in a device, a ranking of how successful the participant thought each device would be for completing the tasks, and an indication of which device the participant would choose to use if they could only pick one. We would later ask the same questions in the post-test questionnaire to discover whether their preferences had changed.

Following the completion of the pre-test questionnaire, we showed participants a map of the marker locations, handed them a single device, and instructed them to move to the starting marker before proceeding to complete the tasks. We suggested tasks be completed sequentially but did not enforce it as a requirement. In addition to collecting data from questionnaires, we also wanted to capture participants' reactions towards the devices as they were being used. We therefore employed the 'think aloud' protocol by asking participants to verbalise what they were thinking as they progressed through the tasks with each device. Comments were digitally audio recorded and we also took supplementary notes on our observations. We decided not to use video recording due to the impracticalities involved with maintaining a usable camera angle on each participant as they moved between tasks. After completing each task, we asked for a verbal score of how easy the task was using a SEQ (Single Ease Question) rating scale (Sauro and Dumas 2009, Sauro 2010) from one to five, where one represented 'very difficult' and five 'very easy'. Once all tasks had been completed with a single device, we asked participants to fill out a SUS (System Usability Scale) questionnaire (Brooke 1996) rating the device's usability. The SUS questions were unmodified save for the replacement of the word *system* with the word *device* and clarification of 'various functions' mentioned in question five to include relevant hardware features such as camera, screen, audio, and so on. We then handed the participant another device and instructed them to repeat the process. The order in which participants used the devices was counterbalanced.

After completing all tasks with all devices, we asked participants to fill out a post-test questionnaire. The post-test questionnaire asked participants to rate how effective they thought each device form factor was for completing the tasks on a five-point Likert scale. We also asked for a rating of any fatigue felt while using the devices on a similar five-point scale. Finally, the same three questions mentioned previously from the pre-test questionnaire were asked (slightly re-worded) to see if there were any differences in responses. Each participant took approximately 60 minutes to complete the experiment.

6 Hardware

The experiment made use of three current (at the time of testing) Android-based handheld devices, each selected to represent the three form factor categories being evaluated. The smartphone was a Nexus 4 made by LG. It provided a 4.7 inch screen (320 pixels per inch [PPI]), had physical dimensions of 133.9 by 68.7 by 9.1 millimetres, and weighed 139 grams. The mini tablet was a Nexus 7 made by ASUS. It provided a seven inch screen (323 PPI), had physical dimensions of 114 by 200 by 8.65 millimetres, and weighed 290 grams. Finally, the tablet was a Nexus 10 made by Samsung. It provided a ten inch screen (300 PPI), had physical dimensions of 263.9 by 177.6 by 8.9 millimetres, and weighed 603 grams. A visual size comparison between all three devices (proportionally correct) is provided in Figure 3. All devices were configured to run Android version 4.2.2.

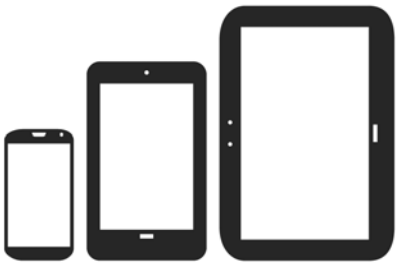

Figure 3: Left: Nexus 4; middle: Nexus 7; right: Nexus 10

7 Results

One of the first questions we asked participants in the pre-test questionnaire was whether they currently owned a handheld device from any of the three form factor categories. All but one of the participants indicated owning a smartphone (93%), one third (67%) a tablet, and only two (13%) a mini tablet. Given this distribution, the mini tablet was by-and-large the form factor with which participants had the least experience.

In addition to surveying which handheld form factors were currently owned by participants, we also asked for a rating of various hardware factors believed to be important for AR use. We asked this question twice, first in the pre-test questionnaire, before any devices were used, and again in the post-test questionnaire. Ratings for each factor were given on a five-point Likert scale that ranged from 'extremely unimportant' to 'extremely important'. We ranked the factors based on the results and present a summary in Table 2 (over page) along with an indication of any position change between pre- and post-test orderings.

The largest position change was between performance and weight, which swapped positions each moving four places. Participants initially considered weight to be a low priority, placing it last, and performance to be a high priority, placing it second. Screen size and camera quality also swapped positions, but only by a single place, while ergonomics and screen quality did not change. Ergonom-

ics can be considered a key aspect of usability and so its place atop the list is not surprising.

Table 2: Pre and post-test ranking of hardware factors

Rank	Pre-test	Post-test	Change
1	Ergonomics	Ergonomics	-
2	Performance	Weight	+ 4
3	Screen size	Camera quality	+ 1
4	Camera quality	Screen size	− 1
5	Screen quality	Screen quality	-
6	Weight	Performance	− 4

In terms of the position swap between performance and weight, we suggest the initial placement of these factors was based on participants' experiences with the way they currently use their device(s). It is reasonable to assume that the majority of device usage would not have involved AR but instead ephemeral in-and-out experiences such as reading an email, browsing the Internet, replying to a text message, checking social media, or playing a quick game. These sorts of everyday examples are likely to place more value on ergonomics, performance, and screen size because they are all factors that have an appreciable effect on these tasks. Furthermore, during normal use, it is unlikely that a device will be held in an orientation/position for long enough that weight will become a significant issue. AR obviously presents different demands. Certainly, given the nature of the tasks developed for this study, we can understand more importance being placed on weight as a factor, but we also believe it to be an important consideration for any handheld AR device.

Similarly, as such a large part of the AR experience on a handheld device relies on the rear-facing camera, we're not surprised to see camera quality swapping positions with screen size. As noted by many of the participant comments and feedback, the difference in screen size did not have as much of an impact on the AR experience as they might have initially thought. It was more important that the images remain responsive and free from 'lag', which most associated with camera quality.

Although performance moved to last place in the post-test ordering, we do not consider this an indication that it is unimportant but rather an acknowledgement of the state of mobile hardware today. As we mentioned in the introduction, handheld devices now come equipped with impressive specifications including extremely powerful CPUs and GPUs. The majority of handheld hardware is now at the very least 'good enough' and all three of the devices we used were recent enough for performance not to be a concern. When the hardware is at this level, it is expected that users will focus on other areas when considering issues.

Overall, we find the post-test ranking to be a fairly accurate ordering of what we would consider important factors for handheld AR usability.

7.1 SUS

To study the usability of device form factors further, we administered the System Usability Scale (SUS) to each participant after they had completed all tasks with each device. Evaluating the SUS questionnaire responses involved calculating a score out of 100 that, in our case, represented the overall usability of the device in question. Each participant completed three SUS questionnaires in total, one for each device they used. While the SUS scores related specifically to the devices tested, we can extrapolate the results to form a more general impression of the form factor categories they represent; we are less interested in the actual scores for each device and more interested in the differences between them.

There are no strict rules regarding what scores correlate to something that is considered 'usable'; however, a rule-of-thumb interpretation given by Bangor, Kortum *et al*. (2008) serves as a useful scale to reference. For our purposes, we consider anything below 50 unacceptable, anything between 50 and 70 marginal, anything above 70 passible, and anything above 85 ideal.

A summary of the overall SUS scores for each form factor is given in Table 3 and a boxplot of the results is presented in Figure 4. Based on these scores, the smartphone achieved a rating of *ideal* while the mini tablet and tablet achieved a rating of *passible*. We point out, however, that the first (81.25) and third (93.75) quartiles for the smartphone and mini tablet were identical and the mini tablet was only 0.5 below an ideal rating (dragged down by its outlier of 37.5). The lower standard deviation for the smartphone does support our observation that, collectively, participants appeared to be more comfortable using it. On several occasions, we noted participants saying the smartphone felt "familiar" and "easy to use".

Table 3: Summary of SUS scores for each form factor

Form Factor	M	SD	Min	Max
S	88.83	7	80	100
MT	84.5	14.49	37.5	97.5
T	77.83	13.19	52.5	100

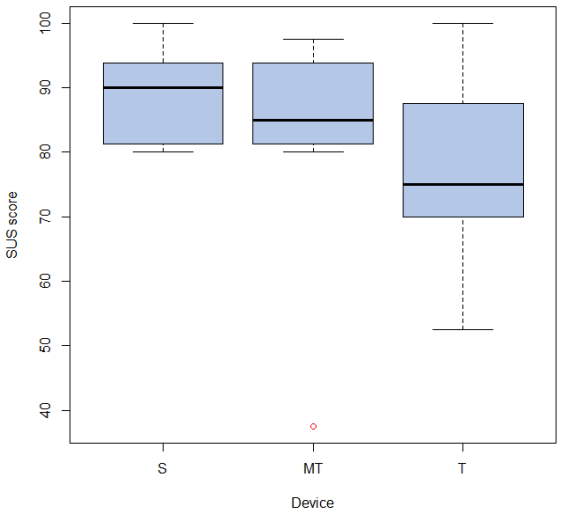

Figure 4: Boxplot of SUS scores for each form factor

To investigate the results further, we ran a repeated measures ANOVA on the SUS scores. Mauchly's Test of Sphericity indicated that sphericity had not been violated ($\chi^2(2) = 1.95, p = 0.378$) and analysis of the ANOVA output revealed a significant difference between the means of the scores for the three form factors

$(F(2, 28) = 4.73, p = 0.017, \eta_{p^2} = 0.252)$. To further investigate the significance between SUS scores for distinct form factor combinations, we ran a Bonferroni post hoc test. The results showed no significant difference between the smartphone and mini tablet ($p = 0.799$) or between the tablet and mini tablet ($p = 0.372$). It is likely the size difference between these combinations was too small to make any discernible difference in the minds of participants. This matches the trend we saw in participant device ownership in terms of the types of devices that were owned. The difference between smartphone and tablet ($p = 0.006$), on the other hand, was found to be significant and shows that a usability difference does exist between these two form factors, which we would expect given the relatively large size gap between the two.

7.2 SEQ

The SUS scores served as a way to measure a device's usability on the whole but didn't shed any light on individual tasks. To get a sense for how well participants felt the form factors performed for each task, we made use of the Single Ease Question (SEQ). The combined results are presented in Figure 5 and show the mean task score for each device-task combination.

Participants rated the mini tablet and tablet ahead of the smartphone for task one suggesting the increased screen size was beneficial for maintaining focus over the target. Participant comments supported this with many remarking that they found the bigger devices "a lot more stable", "easier to hold steady", and "better than I thought". The screen size specifically was noted as providing a better view of the crosshair and virtual objects on screen. We also recorded a few occasions where participants commented that the larger devices didn't appear to accentuate small movements as much as the smartphone; however, given a similar frequency of comments related to lag in the smartphone's camera, we feel this was likely due to the hardware rather than any significant difference.

The task two results show a distinct 'staircase effect' whereby the smartphone comfortably leads the mini tablet which in turn leads the tablet. As this task required the device be held at or above eye level, these results were not surprising. A heavier device was expected to be more difficult and uncomfortable to hold in that orientation and would only have been exacerbated by the need to also interact with the screen. Many participants commented on the difficulties associated with holding the tablet up with one hand while also trying to interact. Terms such as *insecure* and *slipping* were often used to describe the experience.

Both the mini tablet and tablet lead the smartphone in task three with the same score. Given this task was the first to involve a drawing interaction, it was expected that some participants would notice the inherent occlusion issues associated with small screens. Although no one had any problems drawing the path, many made note of their finger/hand blocking their view and a general lack of accuracy, which was more pronounced on a smaller screen.

Task four was presented similarly to task one but differed in that the target continuously moved in a circular orbit. We expected the smartphone to score much better given its weight and manoeuvrability advantages, but the scores suggest screen size is also an important factor for moving targets with the device offering the best compromise between the two (the mini tablet) leading the others quite noticeably. Although we hadn't originally considered it, an obvious effect of using a smaller device for this type of task compared with a larger one is the need to make more dramatic motions to maintain focus. This subsequently leads to increased chances of noticing visual/tracking lag.

Task five was the first of the free-form drawing tasks and presented the drawing surface vertically thereby forcing the device to be held out in front of the participant. The results were similar to task two with weight clearly being the main issue encountered. Many participants commented on the difficulty with which they found holding the tablet up with one hand while trying to draw with the other and frequently discovered the hand holding the device beginning to induce wobble as a result of wrist strain. We observed many different approaches to holding the tablet up in this orientation ranging from gripping it along the left/right bezel, holding it with both hands and drawing with the thumbs, and holding it like an easel from the bottom. While lighter devices didn't suffer from the same levels of wrist strain as the tablet, they instead suffered from their own type of wobble brought about from interactions with the screen—lighter devices were more susceptible to movement in this regard as a result of finger pressure. Many of the issues observed could be addressed by implementing some form of interaction technique whereby the AR view is 'frozen' allowing the device to be repositioned to a more comfortable orientation for drawing while at the same time removing wobble. We discussed examples of such techniques in section 2.

When the drawing surface was oriented horizontally,

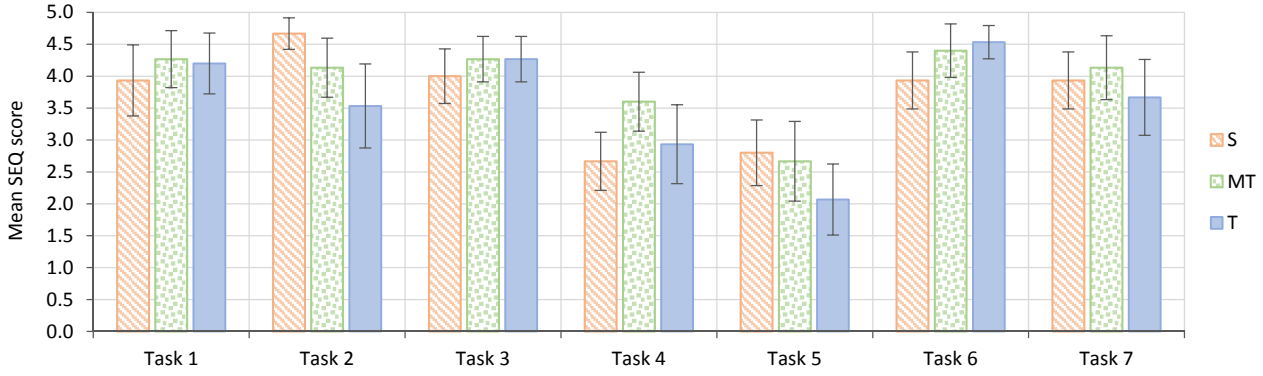

Figure 5: Mean SEQ scores for smartphone (S), mini tablet (MT), and tablet (T)

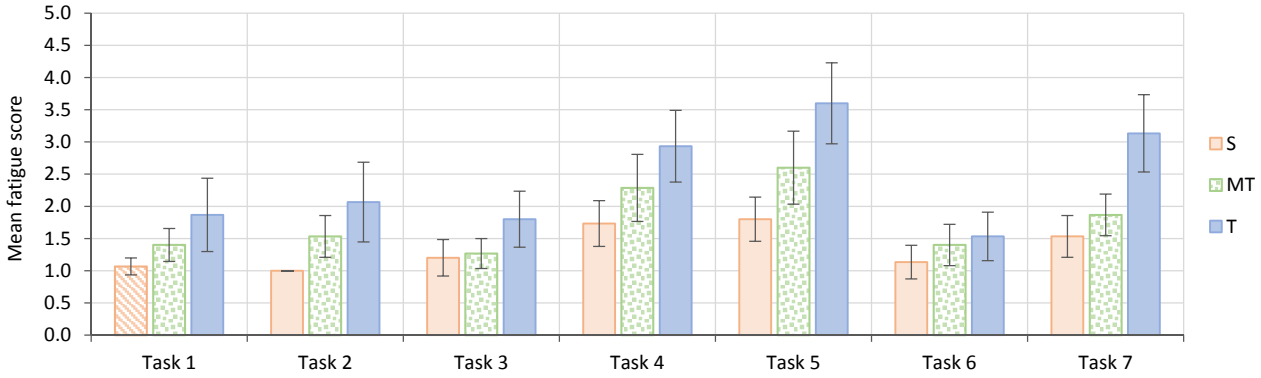

Figure 6: Mean fatigue scores for smartphone (S), mini tablet (MT), and tablet (T)

as in task six, the results are a mirror of those from task five. In this orientation, weight does not overcome the advantages offered by a larger screen and the tablet scores much better. It is clear that drawing in a vertical orientation is undesirable and we frequently recorded participants expressing how the horizontal orientation made the task more enjoyable and easier to complete.

Task seven, again, saw the tablet placing last but interestingly the mini tablet first; we expected the smartphone to place first given the task was prolonged and involved moving between markers. Most participants said they found the task easy but tedious and repetitive. None found it particularly straining even though we required the device be held vertically when carrying water. This suggests that participants didn't mind engaging with an AR task that spanned multiple locations, although one participant did comment that they "felt silly holding a tablet like this [arms stretched out] while walking".

The clear message from the SEQ results is that no one device appears best-suited to all tasks. Certain form factors are more appropriate for certain task types. This was echoed by many of our participants who similarly commented that they felt the choice as to which device they would prefer to use was very much task and location dependent. Smaller devices might be preferred in public settings due to the perception associated with holding up a large device whereas collaborative tasks might view larger devices as an advantage, allowing multiple users to more easily share an experience. The 'right' device can also be influenced by personal ergonomics. What is uncomfortable for one person might be comfortable for another; as one participant said, "I have big hands so for me a tablet is pretty much a smartphone".

7.3 Fatigue

Given the differences between SEQ scores, we wanted to investigate whether device fatigue had any influence. We asked participants in the post-test questionnaire for a rating of any fatigue felt while completing the tasks. The rating was given on a five-point Likert scale ranging from 'no fatigue' to 'extreme fatigue'. The results are presented in Figure 6 and show the mean fatigue score for each device-task combination. The results reflect the ascending weight order of each device; in other words, the lightest device received the lowest fatigue score while the heaviest device received the largest. This was expected.

To see if there was a relationship between fatigue and SEQ rating, we ran a Pearson correlation for each form factor and found that all three form factors exhibited a strong negative correlation between fatigue and SEQ score. The mini tablet ($r = -0.92, p = 0.004$) was found to have the strongest relationship followed by the tablet ($r = -0.9, p = 0.006$) and finally the smartphone ($r = -0.9, p = 0.006$). All results had p-values less than 0.05 and show that fatigue was likely one of the contributors to a participant's SEQ rating. This seems reasonable as we would expect participants to find tasks that were more fatiguing more difficult.

With respect to device ergonomics having an effect on fatigue, some of the comments we recorded during the sessions suggested that a grip or handle of some sort could assist with holding the larger devices. These comments mainly occurred while participants were completing task five ("Some sort of pad on the back perhaps; something I can grip") and seven ("I feel like I want a handle…"). The work of Kruijff and Veas aims to address device ergonomics via the Vesp'R product.

7.4 Overall Preferred Form Factor

Finally, as a subjective measure of form factor preference, we asked participants in both the pre- and post-test questionnaires which single form factor they would choose to use for the tasks if they could only pick one. Answers from both questionnaires are provided in Table 4 and we have highlighted rows where the responses changed. We have also included the form factors rated most usable from each participant's SUS scores and highlighted those where there is a partial or full match with the post-test form factor.

In the pre-test results, the mini tablet was rated most preferable eight times followed by the smartphone five times and the tablet twice. In the post-test results, the mini tablet was rated most preferable nine times with the smartphone staying at five and the tablet dropping to one. Almost half of the participants (46%) changed their response between pre- and post-test. Of those that did, four changed to a smaller form factor and three to a larger. The preferred form factor as indicated on the post-test questionnaire matched the (equal) highest rated form factor in terms of SUS usability 73% of the time.

Table 4: Pre- and post-test preferred device

ID	Pre-test	Post-test	SUS
1	S	MT	S, MT
2	MT	S	MT

ID	Pre-test	Post-test	SUS
3	MT	MT	MT
4	T	MT	S, T
5	MT	MT	S
6	T	MT	MT
7	MT	MT	MT
8	MT	S	S
9	S	S	S, MT
10	MT	T	S, MT
11	MT	MT	S, MT
12	MT	MT	S, MT, T
13	S	S	S
14	S	S	S
15	S	MT	S, MT

8 Conclusions and Future Work

We have presented discussion and results on the usability of different form factors for handheld augmented reality use. Three form factor categories were evaluated: smartphone, mini tablet, and tablet. The distinction between these categories was based on a survey of currently available devices. Testing involved using a device from each category to complete seven tasks, each of which was designed to incorporate typical AR interactions.

We asked participants to provide an SEQ rating following the completion of each task and complete a SUS questionnaire following the completion of all tasks with a particular device. The SEQ scores revealed no one form factor was best-suited to all tasks. The 'right' form factor was found to be influenced by the specific task, to some extent its location, and personal preference. SUS scores revealed the smartphone to be the most usable form factor followed closely by the mini tablet. Further analysis revealed a significant usability difference between smartphone and tablet but no such difference between smartphone and mini tablet or between tablet and mini tablet—in these cases, it was suggested that participants were unlikely to perceive a difference. Fatigue ratings were also provided following each task and reflect the ascending weight order of the devices used. We found a strong negative correlation between device weight and SEQ score. Overall, participants subjectively found the mini tablet to be the most preferable form factor followed by the smartphone and finally the tablet.

Future work will incorporate our results into further investigations on the use of these form factors for authoring AR content in-environment.

9 References

Arth, C. and D. Schmalstieg (2011). "Challenges of Large-Scale Augmented Reality on Smartphones." Graz University of Technology, Graz: 1-4.

Azuma, R. T. (1997). "A survey of augmented reality." Presence **6**(4): 355-385.

Bai, H., G. A. Lee and M. Billinghurst (2012). Freeze view touch and finger gesture based interaction methods for handheld augmented reality interfaces. Proceedings of the 27th Conference on Image and Vision Computing New Zealand. Dunedin, New Zealand, ACM**:** 126-131.

Bangor, A., P. T. Kortum and J. T. Miller (2008). "An empirical evaluation of the system usability scale." Intl. Journal of Human–Computer Interaction **24**(6): 574-594.

Brooke, J. (1996). "SUS: A quick and dirty usability scale." Usability evaluation in industry **189**: 194.

Dey, A., G. Jarvis, C. Sandor and G. Reitmayr (2012). Tablet versus phone: Depth perception in handheld augmented reality. Mixed and Augmented Reality (ISMAR), 2012 IEEE International Symposium on.

Güven, S., S. Feiner and O. Oda (2006). Mobile augmented reality interaction techniques for authoring situated media on-site. Proceedings of the 5th IEEE and ACM International Symposium on Mixed and Augmented Reality, IEEE Computer Society**:** 235-236.

Hoober, S. (2013). "How Do Users Really Hold Mobile Devices?" Retrieved 9 July, 2014, from http://www.uxmatters.com/mt/archives/2013/02/how-do-users-really-hold-mobile-devices.php.

Kato, H. and M. Billinghurst (1999). Marker tracking and HMD calibration for a video-based augmented reality conferencing system. Augmented Reality, 1999. (IWAR '99) Proceedings. 2nd IEEE and ACM International Workshop on.

Kruijff, E. and E. Veas (2007). Vesp'R - Transforming Handheld Augmented Reality. Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, IEEE Computer Society**:** 1-2.

Kruijff, E. and E. Veas (2008). Vesp'R: design and evaluation of a handheld AR device. Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality, IEEE Computer Society**:** 43-52.

Kurkovsky, S., R. Koshy, V. Novak and P. Szul (2012). Current issues in handheld augmented reality. Communications and Information Technology (ICCIT), 2012 International Conference on.

Langlotz, T., S. Mooslechner, S. Zollmann, C. Degendorfer, G. Reitmayr and D. Schmalstieg (2012). "Sketching up the world: in situ authoring for mobile Augmented Reality." Personal and Ubiquitous Computing **16**(6): 623-630.

Lee, G. A., U. Yang, Y. Kim, D. Jo, K.-H. Kim, J. H. Kim and J. S. Choi (2009). Freeze-Set-Go interaction method for handheld mobile augmented reality environments. Proceedings of the 16th ACM Symposium on Virtual Reality Software and Technology. Kyoto, Japan, ACM**:** 143-146.

Sauro, J. (2010, 2 March). "If You Could Only Ask One Question, Use This One." Retrieved 22 July, 2014, from http://www.measuringusability.com/blog/single-question.php.

Sauro, J. and J. S. Dumas (2009). Comparison of three one-question, post-task usability questionnaires. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM.

Vincent, T., L. Nigay and T. Kurata (2013). Precise Pointing Techniques for Handheld Augmented Reality. Human-Computer Interaction – INTERACT 2013. P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson and M. Winckler, Springer Berlin Heidelberg. **8117:** 122-139.

Wagner, D., T. Pintaric, F. Ledermann and D. Schmalstieg (2005). Towards massively multi-user augmented reality on handheld devices, Springer.

Wither, J., S. DiVerdi and T. Hollerer (2007). Evaluating Display Types for AR Selection and Annotation. Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on.