This work is protected by copyright and other intellectual property rights and duplication or sale of all or part is not permitted, except that material may be duplicated by you for research, private study, criticism/review or educational purposes. Electronic or print copies are for your own personal, non-commercial use and shall not be passed to any other individual. No quotation may be published without proper acknowledgement. For any other use, or to quote extensively from the work, permission must be obtained from the copyright holder/s.

The usability of knowledge based authentication methods on mobile devices

James Rooney

PhD in Computer Science

Keele University

December 2013

Abstract

Background

Mobile devices are providing ever increasing functionality to users, and the risks associated with applications storing personal details are high. Graphical authentication methods have been shown to provide better security in terms of password space than traditional approaches, as well as being more memorable. The usability of any system is important since an unusable system will often be avoided.

Aims

This thesis aims to investigate graphical authentication methods based on recall, cued recall and recognition memory in terms of their usability and security.

Method

A Systematic Mapping Study and two experiments were performed. Criteria were developed to help evaluate and compare the models in terms of their usability and security. The first experiment compared three different types of graphical authentication method, represented by Awase-E, DrawASecret and PassPoints. Two groups of participants were used, a set of IT literate students and a set of less technologically aware members of the public. The second experiment compared two variations of the Awase-E method, both improving the security of the method in different ways, using another student group.

Results

In the first experiment, only small differences were found between the usability scores of the methods, although Awase-E was shown to be the most usable of the three. The second experiment showed that the participants preferred remembering just four ordered images

rather than having to remember five unordered images mostly due to the requirement to associate the images together.

Conclusion

The criteria have been shown to be able to evaluate the three authentication methods discussed in this thesis, and could be adapted to evaluate other methods. Recognition methods have generally been shown to be more usable but at the cost of security, and so more work needs to be done into this area to find ways of reducing the security risk.

Acknowledgements

I would first of all like to acknowledge and thank the EPSRC who provided the funding for the PhD as a whole.

I would like to thank Pearl Brereton, my supervisor, for all of the support and advice she has given me over the entire course of my PhD. I would also like to thank Mark Turner, who was my supervisor for the first two years of the PhD and helped to start me off on my work. Also I would like to thank Barbara Kitchenham for her invaluable support as my second supervisor, particularly for her help and advice with the statistical side of my research. I would also like to thank all of the participants who took part in the experiments I undertook, without whom this research would not have been possible.

Finally I would like to thank my parents, not least for their help in proof reading my work, but also for being there to support me through the whole process.

Contents

List of Tables

List of Figures

1 Chapter One – Introduction

This chapter covers the main research objectives being investigated, which focus on how to design and evaluate usable security systems for mobile devices. Because of the small size of such devices, which are continually being made smaller and more complicated, a full security system which would be appropriate for a desktop or laptop is not practical. The thesis therefore explores the application of security systems which can be more easily used on small mobile devices and evaluates the methods currently available for this purpose. Mobile devices are no longer devices purchased entirely for their functionality, but for the pleasure and ease of use, and any security system used must also be acceptable to the user.

1.1 Motivation

Over the last few years, mobile devices have become ever more complex, leading to the current set of smartphones, capable of many functions beyond simple telephoning and texting. Modern smartphones, such as Google Android based ones, or the iPhone, are capable of web browsing, storing documents and even sensitive functions such as accessing bank accounts, and, more recently, Near Field Communication (NFC) payments in shops. With the ever increasing functionality of such devices comes a greater risk of data/security breaches if the devices are lost or stolen, which was less of an issue with more traditional desktop systems. There have recently been many high profile examples of laptops being lost with extremely sensitive data stored on them (*Programming stuff*, 2007, Calin, 2009, Carey, 2012, Kelly, 2012), and as more people begin to use smartphones and tablets for work purposes, this is likely to increase.

Increasing numbers of services do not require any data to be stored on the device itself, the device simply acts as a gateway for the application or website, allowing the user access to the information resource. This makes the authentication step even more important than before since it is the only way to access the information stored or required and a weak authentication procedure cannot be balanced by, for example, encrypting data stored on the device in case of theft.

One of the quickest and most basic ways to secure a phone through the use of an authentication method or strategy is through the use of a PIN number, and this feature has been available on mobile phones for many years. By convention these PIN numbers almost always use four digits, although there is no particular reason this could not be increased, particularly with modern smartphones. This is in itself not a very secure method since there are only 10,000 (10^4) possible combinations of 4 digit numbers available (although if the mobile is locked after a set number of these, the risk is significantly decreased). Beyond the basic PIN, Google android phones currently have access to two alternative methods for unlocking the phone. The first uses a standard password unlock using alphanumeric characters and the second using a pattern unlock method performed by joining up a series of nodes on the screen into a pattern.

Research has shown that people are better at remembering images than they are at memorising strings of text or numbers (Shepard, 1967, Tulving & Watkins, 1973, Hollingworth, 1913). As such, many different graphically based authentication methods have been proposed, originating with Blonder's method (Blonder, G.E., 1996), a full description of which is given in Section 3.1.2.1. Many of the new methods proposed are intended to highlight or fix an issue with an existing method, or to improve them in one

particular way, but are not put through a rigorous usability evaluation to see if people would actually like or be able to use them. A selection of Graphical Authentication Methods is discussed in Chapter Three.

Usability research has also been around for a long time, with a substantial body of work about how to design usable interfaces for the web, such as the Nielsen Usability Heuristics (Nielsen, 2005) and the Usability.gov guidelines (HHS Web Communications Division, 2009). The subject of Usability as a whole is discussed in detail in the literature review into it in Chapter Four. Many of the concepts and guidelines already developed can be applied in some way to mobile devices, or adapted to face the specific challenges associated with them (such as the substantially decreased screen size available for imparting information). The interface to the system is also very different from that of a desktop in that there is a small numeric keypad on older phones and a touchscreen 'virtual' keypad on newer smart phones. The usability of these data entry mechanisms for the system also needs to be evaluated as to whether it is good, or appropriate, enough.

Usability is subjective, so when a new system is designed, there has to be input from the users as to what they consider usable, rather than simply what the designer(s) believe. Furthermore, usability studies will sometimes be performed on groups of students within the research institute which has produced the new method being evaluated. This can lead to biased results, particularly with technological studies, which would not be applicable to the population as a whole, who are likely to have less technological awareness than the students.

The focus of this work is to compare different types of graphical authentication methods to discover which is the most usable. The methods chosen for the evaluation were chosen as representative samples of the three main types of graphical based authentication from the larger sample discussed in Chapter Three.

Once the most usable of the three methods has been discovered, a new method based on the method previously found to be the best can be created, and again usability tested to show what effect the change has on the usability of the method. As a result of this research, recommendations will be made about what future work on graphical authentication methods should be undertaken to ensure that people will be willing to use such methods in the long term.

1.2 Research Objectives

The overall aim of this thesis is to investigate whether graphical authentication methods are both usable and secure on mobile devices. This breaks down into three objectives:

1. Develop criteria to evaluate whether a graphical authentication method is both usable and secure.
2. Use the criteria to determine experimentally which type of graphical authentication method is the most usable.
3. Extend one of the existing methods, building on the results of the experiment, to improve either the usability or security aspect of the method.

1.3 Contribution to Knowledge

This work contributes to the knowledge in the research area in several ways. It is the first of its kind to empirically compare different types of graphical authentication method. Previously, comparisons have often been based strictly on one aspect of a new method to show its advantages over an older one, rather than a more general usability comparison of the two. To do this, a set of criteria was developed which was used to evaluate the different methods in terms of both security and usability. However, the criteria would be easily applicable to other mobile authentication methods, or even other mobile applications, to help determine whether they are usable.

In order to evaluate the usability of the methods, a set of technologically confident students and a set of participants with less experience of technology, whether mobile or otherwise, were used. Many studies use only a student group of participants as they are readily available within a university setting. This however can introduce a bias into the results since these students are likely to be familiar with mobile devices, and so the results might not be applicable to the more general population.

From the experiment, it is possible to more reliably indicate whether or not the methods are likely to be usable by the population as a whole, and whether those people would actually want to use the methods in the first place.

1.4 Thesis Outline

The thesis is organised as follows. Chapter Two details the process of carrying out a Systematic Mapping Study into mobile security as a whole along with the discussion of its

results. The results of this show that there is a lack of work into the Usability side of mobile security methods. From these results, more work has been done into the area of authentication rather than other security elements. This thesis, therefore, focuses on the usability of authentication methods for mobile devices.

Chapter Three looks into authentication methods and concepts on mobile devices as a whole where authentication is used in terms of a knowledge based login method. It discusses the three aspects of mobile authentication (something a user has, knows or is) as well as looking in more detail at graphical authentication methods.

Chapter Four provides a general overview of usability, beginning with formal definitions of usability from a range of sources and moving on to how it is implemented and used in a variety of situations. It also discusses some of the psychological aspects of usability along with the different types of memory people use to perform tasks, and their impact on the usability of a system.

Chapter Five details an experiment (Experiment 1) into the security and usability of three of the currently available methods for mobile authentication, Awase-E, DrawASecret and PassPoints. The 11 criteria used to evaluate them are also described along with their derivation from currently existing guidelines into usability and security. Implementation details relevant to the specific versions of the three methods created for the Android operating system are also discussed.

Chapter Six discusses the results of Experiment 1. It shows how Awase-E was the preferred, and most usable, of the three methods and discusses the implications of this in terms of security.

Chapter Seven discusses ways of extending the original three methods to improve either their security or usability. Two of the variations improving the security of AwaseE are chosen to be the subject of a further experiment (Experiment 2) to determine what effect the additional security features have on the usability of the original method. The experimental process for Experiment 2 is also discussed, although briefly, as it is strongly based on the method used in Experiment 1.

Chapter Eight discusses the results and implications of Experiment 2, and shows that the variation based on ordering the images the user had to chose was more useable than the simply requiring the user to remember more images.

Chapter Nine discusses the results of the experiments in terms of the overall thesis, as well as providing a summary of the main points of interest. Finally Chapter Ten presents the conclusions of the work as well as detailing potential further work into the area.

2 Chapter Two – Mobile Security

This chapter details the Systematic Mapping Study which was performed in the area of Mobile Security in order to identify the relevant work within the field. The study shows that the most work within the field has been undertaken in the area of identity establishment which is often referred to as authentication (Benantar, 2006). The study also highlights a lack of research into how usable the developed security systems were in practice.

2.1 Background

Recently, a significant amount of research has been published in different areas within the field of mobile security. A Systematic Mapping Study was proposed as the best method to identify research clusters as well as any potential gaps in the research area (Kitchenham & Charters, 2007). The study enabled the collation and categorisation of all of the available information in this area, with the intention of making it easier to see the various subtopics as well as showing where current research was being concentrated.

> *"Systematic Mapping Studies (also known as Scoping Studies) are designed to provide a wide overview of a research area, to establish if research evidence exists on a topic and provide an indication of the quantity of the evidence." (Kitchenham & Charters, 2007)*

This section details the process of performing the Systematic Mapping Study.

2.2 Protocol

A protocol is used in a Systematic Mapping Study to define how the study will be carried out. The search strings to be used, along with the inclusion and exclusion criteria are

defined in the protocol before the study takes place so that there is a consistent structure to enable a thorough search to be performed.

2.2.1 Research Questions

The systematic mapping study had the following main research questions:

RQ1) Which journals and conferences are publishing work relating to security on mobile devices?

RQ2) What elements of security are inspected?

RQ3) What types of study have been performed?

Studies will be classified using categories as defined by Tichy et al. (1995) as follows.

1) Theory: A study in which only a theoretical model is proposed, no suggestion for implementation is provided.
2) Design & modelling: Articles in which a working model is presented for some purpose, but has not necessarily been evaluated to indicate their validity.
3) Empirical: Studies whereby a known design or method is tested to see whether it achieves the claims made of it.
4) Hypothesis testing: Articles in which a formal hypothesis is presented along with a method of testing it.
5) Other: Any study that does not fall into one of the above categories.

Security itself can be split into several elements, as defined by Benantar (2006) below:

1) Identity establishment: This is the way in which a user or process is confirmed to be who they say they are. This is also known as authentication.
2) Access control: Controlling access permission policies for users or devices. This is often referred to as authorisation. Once a user, or process, has authenticated itself

on a system, they are then granted access to whichever areas of the system they have permission to access.

3) Availability: Ensuring that resources either on a device or on a network are available, such as by preventing or minimising the effect of denial of service style attacks
4) Data and message security: Protection of data both stored on a device and transmitted to or from it. This is most often achieved through the use of encryption.
5) Non-repudiation: Preventing a user or device from falsely denying its part in an activity. This is often realised through the use of audit and transaction trails to record the actions of individual users or entities.

Some works however will fit into all of these elements as they are concerned with developing a whole security system in one way or another that would address all concerns. Such models or systems can be considered as a whole Security Suite.

2.2.2 Search Strategy

The search strategy used for the Mapping Study was as follows:

a) Identify major terms and keywords relating to mobile security

b) Identify alternative spellings and synonyms for keywords and phrases

c) Use the Boolean OR to identify alternative spellings or synonyms

d) Use the Boolean AND to link the major terms and keywords

The term "mobile availability" was not included in the search string since preliminary piloting of terms indicated that this would have returned too many irrelevant publications.

Following this strategy, and using the results from the piloting activities, the following search string was derived:

(mobile OR "mobile device" OR "mobile devices") AND (security OR "access control" OR authorisation OR authorization OR authentication OR nonrepudiation OR non-repudiation OR "non repudiation" OR "data authenticity" OR "encryption")

Validation of the search string was carried out by ensuring the following papers appeared in the results:

Management of security policies for mobile devices (Green, 2007)

Security of Mobile Agent-Based Web Applications (Ou et al., 2008)

An Investigation into Access Control for Mobile Device (Perelson & Botha, 2004)

Access control for future mobile devices (Chen & Sivakumar, 2005)

These papers were known papers which could be used as validation papers for the search strings to ensure relevant papers were found. Although the Access Control element was not the focus of the Mapping Study, it is one of the elements of Security as a whole and so these four papers should appear in any search relating to Mobile Security.

The following digital libraries were used to search for material for the study

ACM Digital Library (http://portal.acm.org)

IEEExplore (http://ieeexplore.ieee.org/)

CiteSeerX (http://citeseerx.ist.psu.edu/)

The IEEE Xplore and ACM digital libraries are the main repositories for computer science research material online, whereas CiteSeerX indexes a number of other resources, such as

Science Direct. Web of Science was not included in the Mapping Study since the results found were felt to be a sufficient sample of the available research.

2.2.3 Selection Criteria

The inclusion and exclusion criteria were designed to filter out irrelevant results from the set of results returned by the search strings.

Inclusion Criteria

- Only peer reviewed literature was included, such as journal articles and from conference proceedings.
- Literature relating to problems and solutions to security systems on mobile devices.
- Literature relating to the security of data on mobile devices, whether transmitted or physically stored on the device.

Exclusion Criteria

- Literature not related to security on mobile devices (for example, a paper on mobile/movable security robots)
- Literature in slideshow format
- Literature solely relating to user's views of security systems, or not tackling a problem related to the mobile devices themselves.
- Information that has been, or could have been, edited after its original publication by a person other than the original author (e.g. Wikipedia).
- Grey literature (including textbooks, technical reports, non peer reviewed conference papers)

2.2.4 Data Extraction

The following pieces of information were recorded about each of the publications found from the data sources.

Title

Author

Year of publication

Publication type

Publication location (e.g. journal or conference name)

The element of security investigated

The type of study performed

A brief overview of the paper

Any useful or related references

The papers found were categorised in three ways after the inclusion and exclusion criteria had been applied. The papers were classified as to whether they were from conferences or journal sources. The papers were then further classified by what security element they dealt with, and by the type of study being performed. A sample of the categorised papers was given to two other researchers to verify the classifications made over the course of the study.

Papers which dealt with more than one element were categorised into each of the categories they covered, with papers covering all of the categories being separated into the separate security suite category. If a study fell into more than one element of security, it was counted for each of the elements it fell under. If more than one study was presented in a paper then each study was treated separately. Each of the papers was also classified into

the type of study performed, as defined by Tichy et al. (1995), with the papers containing empirical studies classified further into what type of empirical study performed, as defined by Budgen (2008).

The study was limited to a maximum of 150 papers from each of the online resources, before the inclusion and exclusion criteria were applied, due to the large volume of results returned from the search string.

2.3 RQ1 Results

In total 214 papers were included in the study. Table 1 shows how many of the included papers were from conference proceedings or journals. The majority of papers found are from conference sources rather than journal publications which may be due to the quicker turnaround time for publishing work at conferences rather than as journal articles. The field itself is also relatively new (almost all of the papers were published after 2000) so there may not have been time for work to get through to journal articles, which can take a long time to be published.

Publication Type	Number of Publications
Conference	180
Journal	34

Table 1: Papers classified by publication type

Table 2 shows the publications or conferences which published the most works found during the study. This enabled the researchers to see if there was a location where a high volume of work being published which could be later included in a follow up study to ensure no papers were missed from the search results. For brevity, only conferences or journals with two or more publications are shown.

Type of publication	Name of location	Number of publications
Conference	IEEE Vehicular Technology Conference	7
Conference	IEEE Wireless Communications and Networking Conference	6
Conference	International Conference on Advanced Communication Technology	5
Conference	Hawaii International Conference On System Sciences	5
Conference	International Conference on Availability, Reliability and Security	4
Conference	IEEE International Conference on Communications	4
Conference	IEEE Conference on Wireless and Mobile Computing, Networking and Communications	4
Journal	IEEE Transactions on Wireless Communications	4
Conference	IEEE International Conference on Performance, Computing and Communications	3
Conference	IEEE International Symposium on Personal, Indoor and Mobile Radio Communications	3
Journal	Wireless Personal Communications	
Conference	ACM Symposium on Access Control Models and Technologies	2
Conference	International Symposium on Biometrics and Security Technologies	2
Conference	IEEE Consumer Communications and Networking Conference	2
Conference	International Symposium on Consumer Electronics	2
Conference	Proceedings of the ACM/IEEE Design Automation Conference	2
Conference	Canadian Conference on Electrical and Computer Engineering	2
Conference	International Conference on Information Networking	2
Conference	Information Security South Africa Conference	2
Conference	IEEE Conference on Local Computer Networks	2
Conference	ACM international symposium on Mobile ad hoc networking and computing	2
Conference	International Conference on Mobile Ad-hoc and Sensor Networks	2
Conference	International Conference on Mobile and Ubiquitous Systems: Networking & Services	2
Conference	International Conference on Mobile Business	2
Conference	International conference on Mobile systems, applications and services	2
Conference	International Conference on Parallel and Distributed Systems	2
Conference	IEEE International Conference on Pervasive Computing and Communications Workshops	2
Conference	International Conference on Security and Privacy in Communications Networks	2
Conference	International Conference on Ubiquitous Computing	2

Journal	IEEE Communications	2
Journal	IEEE Personal Communications	2
Journal	IEEE Transactions on Consumer Electronics	2
Journal	International Journal of Network Security	2

Table 2: The conferences and journal locations with the highest frequency of papers found during the Systematic Mapping Study

2.4 RQ2 Results

The distribution of papers across the different elements of security is shown in Table 3.

Security Element	Number of Publications
Identity establishment	114
Data & message security	50
Access control	33
Security Suite	20
Availability	15
Other/Survey	7
Non-repudiation	4

Table 3: Papers found from the Systematic Mapping Study broken down by security element investigated

The total number of papers listed is larger than the number of papers included in the study because some of the papers investigated more than one element of security and so these would be counted twice.

The results show that research into mobile security is concentrated on the identity establishment element particularly. Each of the security elements can be further broken down into sub categories as shown in the following sections, along with an overview of the work relating to them found during the study. Further research on authentication, specifically relating to mobile authentication can be found later in Chapter Three.

2.4.1 Identity Establishment

The Systematic Mapping Study found that identity establishment (also referred to as authentication) was the most studied element. This may partly be because it is the most publicly viewable element of security, with all computer users having a password for at least one account or email address, and so this is the one which would attract the most research.

Authentication itself can be broken down into different subsets based on the type of system to which the authentication model is being applied (wireless networks, MANETs etc) and in a broader sense into three types of information requested in order to authenticate (Wood, 1977):

- Something a user knows (knowledge based authentication, passwords)
- Something a user has (token based authentication)
- Something a user is (biometric authentication)

'Something a user knows' is often found these days in the form of text based or numerical passwords, but also applies to more graphical methods of authentication where the user is required to reproduce a drawing to confirm their identity. Of the papers which dealt with authenticating a user, rather than a device (e.g. onto a mobile network), 21/54 of them looked into a knowledge based authentication method. The majority of these worked under the assumption that a username and password would be provided whilst investigating a different aspect of the authentication procedure.

52 of the papers did not specify exactly how the user's credentials would be gathered by the system, and were concerned with how a device could be authenticated on a network

instead and so gave no details on what the user would have to do in order to use the system.

Most knowledge based systems are in the form of an alphanumeric username/password combination. Other possibilities are available particularly in relation to mobile devices, such as requiring the user to shake or move the device in a predetermined pattern to authenticate the user (Mayrhofer & Gellersen, 2007).

A substantial body of work has been done into studying the security of password systems, with many concluding that the main drawback is a lack of human ability to remember complex enough passwords to be cryptographically secure (Inglesant & Sasse, 2010, Hamilton et al., 2007, Suo et al., 2005, Brown et al., 2004, Yan et al., 2004, Zviran & Haga, 1999). Many people will use the same password for all of their accounts, should they have more than one, or use very simple easy to guess passwords. Very basic social engineering can be used to attack and crack passwords, for example simply attempting the names or birthdays of family members or pets. Two of the papers found did however investigate how secure a system would be if only certain levels of password were used, from insecure ones to merely gain basic access onto the device up to highly secure ones intended to protect bank accounts (Chen & Sivakumar, 2005, Chen & Sivakumar, 2005).

Of the papers included, eight looked at the idea of a Single Sign On procedure, although what was defined as Single Sign On varied. Four of the papers defined it as the ability for a user on a mobile device to sign in once to the device, and be granted access to different services, or websites available (Jeong et al., 2004, Chen & Sivakumar, 2005, Wangensteen et al., 2006, van Thanh et al., 2008), either by use of a standard login system or an external

authentication card/SIM card. Other papers proposed single sign on systems that allowed a user to sign into a network once, and afterwards the user would remain authenticated on the network even if they were to disconnect and reconnect (Cheng et al., 2004, Me et al., 2006).

Responses to the problem of too many passwords often require users to change their passwords over a set period of time, as well as to require them to produce passwords of a certain complexity level or length (e.g. no proper nouns, must include numbers). However this can significantly increase the load on a user's memory, particularly if they are required to remember several such different passwords, leading to the user either forgetting or writing it down somewhere easy to read near to the computer. This voids the point of having a secure password since an attacker would have easy access to the passwords if they simply sat at the workstation of the user.

Work is being done into the field of graphical passwords. Studies have shown that it is far easier for people to remember graphical information, such as pictures or colours, rather than strings of letters or numbers (Rock & Engelstein, 1959, Shepard, 1967, Blonder, G.E., 1996). A more detailed look into recall and recognition can be found in Section 4.3.1. Various graphical methods have been created for this purpose. For instance the DrawASecret method allows users to draw a pattern as their password. This has been partly implemented in the current version of Google Android as a more secure unlock screen, as an alternative to a standard pin. Awase-E uses the method of picture recognition by remembering a previously selected set of pictures from a larger set to login with. PassPoints allows a user to choose a set of memorable locations on an image to allow them to log on. Passfaces is similar to Awase-E in that it involves recognising a series of

pictures, although in this case specifically people's faces, where the human brain is good at remembering small details. Another method of PassPoints created a pass area in a larger image of many small icons. A click anywhere within the area or triangle created by the three previously selected icons would constitute a correct login. Many of these are not suitable, or even possible in some cases, for the small screens associated with mobile devices.

'Something a user has' is often found in the use of identity cards to authenticate a user for access to buildings, but Near Field Communication (NFC) devices are increasingly being used for instant payment systems such as the Oyster Card or mobile payment systems such as Google Wallet. Other methods however include wearable devices which only work once the user (and so also the device) is in proximity to a main workstation, instructing the workstation to unlock and allow the user to work. This overall method has a disadvantage in that it can be easy to lose the pass card or similar device, or have it stolen, and so the user can be relatively easily impersonated. As such, in secure environments, this would often be combined with some form of biometric identification to ensure the card does in fact belong to the user attempting to gain access.

For instance, NFC devices can be used to perform multiple functions such as authentication of users, mobile payments and tracking goods. 18 of the papers found in the Identity Establishment category looked into using NFC devices in some way for authentication. These devices typically have a range of a few centimetres and can allow the transmission of data from device to device within this radius. A specific example of an NFC device is a Radio Frequency Identification (RFID) card. These are small cards which can be fitted to some mobile devices to transmit and receive radio frequencies up to a range

of a few meters. This can allow people, or other devices, with such cards to be automatically authenticated onto the device they wish to use without having to specifically login or perform a more time consuming authentication procedure (Park et al., 2006, Kim et al., 2006, Kim & Kim, 2006, Ikram et al., 2008). Due to the high number of such cards and types of device, different methods have been looked at to produce NFC cards that can emulate many types of other cards. This would allow a mobile user access to many more types of network than would normally be available if they were restricted to a particular card (Park et al., 2005, Madlmayr, 2008, Madlmayr et al., 2008).

One application of this can be found in smart tokens for mobile devices. A user can wear such a token, or carry it around with them, and once the token is in range of the mobile device, the device is unlocked for the user. Out of range the device is completely locked, so if it were to be stolen, it couldn't be used in any way (Al-muhtadi et al., 2002, Decker et al., 2004) This kind of device would be particularly user friendly if implemented properly since it would remove all necessity for the user to authenticate themselves on a device in order to use it, although it should be noticed there is nothing to stop an attacker using it to gain access if the token was lost or stolen within the building itself.

Authentication of mobile devices onto mobile ad hoc networks (MANETs) was found to be a major theme within the mapping study with 19 of the identity establishment papers investigating ways of doing this. These varied from using predetermined security keys or tokens on the mobile device to using the SIM card to gain authentication details from the device's home network as well as trust based methods dependent on other devices on the network (Wang et al., 2007, Sun et al., 2008).

'Something a user is' is based on biometric forms of identification, such as fingerprints or face and retina scans. Fingerprint scanners are already found on some modern laptops (but few if any mobile phones at the time of writing), so inclusion in a mobile device would involve a significant amount of hardware investment. Four of the papers from the mapping study directly looked into the use of biometric information on mobile devices. Of these, the most recognisable method of use is to require the user to scan their fingerprint in as an alternate way of logging onto the computer initially rather than requiring them to type their password in Zheng et al. (2008), but methods such as speech, facial scans (Hazen et al., 2006) and even the gait of the user can potentially be used Derawi et al. (2010). Biometrics represents a far higher level of authentication security compared with a password, which can be forgotten, written down and found by an unauthorised individual, or be set to be something too insecure. This ought to increase the usability of the system since users would not have to deal with potentially complex passwords in order to access information.

One final point to note though is that if the database storing the biometric information is compromised, then the data is no longer secure permanently as there is no way to change characteristics such as fingerprints.

A Trusted Computing platform can be used in conjunction with biometric scanners to produce a more secure mobile device (Zheng et al., 2005, Kuntze & Schmidt, 2006, Leung & Mitchell, 2007). Trusted Computing (Trusted Computing Group, 2005) is a method of computer security whereby only known patterns of behaviour of a device are accepted as legitimate, and this behaviour pattern is programmed in down to the hardware level to ensure that any deviation from the normal can be detected and blocked. Biometric

scanners can also be used as the initial part of an authentication procedure only to produce a unique or anonymous ID for a mobile device connecting to a network (Podio, 2002, He et al., 2004).

Another method for authenticating users is by using location analysis, although this more often applies to the field of authorisation or access control. The location of the user at a given time would determine whether or not they would help determine whether a user should be authenticated onto the system at all (Sastry et al., 2003, Mayrhofer, 2006).

2.4.2 Data and Message Security

Data and Message Security is concerned with the physical security of data stored or transmitted. Often this is achieved through the use of encryption software, whether it is of the data stored on the device itself or in terms of the data transmitted over a wireless connection, or often both.

Data stored on a device is particularly at risk if the device is lost or stolen. Even if the authentication system prevents a malicious user from getting onto the device itself, the storage device could be removed and attached to another machine. Were this to happen and the data not be protected, then anything stored on the device would be instantly available to whoever had found it (Dietiker, 2008).

This can be avoided if a separate server were to act as a remote fileserver for the device. Any files the user wished to create would be transmitted securely to and from the fileserver as and when required (Noponen & Karppinen, 2008). This does introduce the obvious issues of ensuring the data transmitted is properly encrypted, as well as the problem that if

the fileserver is compromised, then a significant number of other users would have their files at risk.

Transmitted data is open to attack from eavesdroppers. Due to the nature of wireless networks, all data is transmitted in all directions rather than directly to the receiving access point or mobile tower. This means that any person with a suitable mobile device in range can in theory receive all the data transmitted from a device. If this is unencrypted then such a malicious user would be able to see all of the information being sent from the device (Wang et al., 2007).

Twelve papers looked at using a public key cryptography solution to improve the security of transmitted data on mobile devices. In this, a server generates a public key which the mobile device can use to encrypt transmitted data. The server then uses its own private key to decrypt received data. For example, Hossain et al. (2008) propose this as a method of increasing the security of SMS messages.

One limitation on secure transmissions on mobile devices is the lower processing power associated with them. This means that encryption methods take longer to perform slowing down the attempted operation. This is often avoided by allowing the server to deal with the more computationally expensive side of the encryption procedure leaving the mobile device only needing to use a less intensive method to transmit data. Lower power cryptography methods can also be used to reduce the load on both the device and the server, and proper authentication by confirming that both parties are validated (Fun et al., 2008). Bin Nafey & Ramanaiah (2008) propose a similar method of using smart cards to

reduce the computational load on a mobile device through the use of the Rijndael algorithm (Jamil, 2004)

Another way to handle this transmission securely is through the use of an external key which would be required for all transmissions with a server. This key would also act as the encryption mechanism for data transferring between the two again allowing the heavy processing to be moved away from the mobile device (Pan et al., 2008).

A hardware solution whereby a separate chip is installed on the device would also allow the encryption process to be moved away from the rest of the device's processing abilities. These chips could be designed and optimised specifically for performing such calculations, but would most likely need to be installed on manufacturing into each device (Arora et al., 2007). They also cite results from some usability testing which show that over 50% of mobile appliance users consider security to be one of the biggest hurdles in adopting new mobile applications. However, given the rapidly evolving nature of the mobile industry, this perception is likely to have altered by now.

2.4.3 Access Control

Access control is the method whereby once a user has been authenticated on a system, the resources of the system are either allowed or denied to them. This will often take the form of an access control list, on which is recorded a list of the users on the system along with a list of the commands and functions they are allowed to execute. This particular method is in use on Unix/Linux operating systems to control access to resources on a computer. This method would not be possible to be implemented in an ad hoc environment where there is no centralised location on which to store the access control list and rules.

Eight of the papers found investigated the possibility of using the user's location for Access Control to resources on a network. Using triangulation between three different wireless transmitters, it is possibly for the central system to determine the location of a mobile user on the network. It can then allow or deny privileges to that user based on their location within the building (Cho & Bao, 2006). Using this method, a company could deny access to all resources from outside of the company's offices, stopping anybody who may have been able to steal a login from using it once they leave the building.

Several other methods exist which place the security system between the user and the main system, meaning any action performed has to be passed through the security system before it can act on the main system. For example, Lee et al. (2008) use a Virtual Machine Monitor (VMM) to control access to the mobile device for the purposes of access control. This reduces the performance overhead of the device itself since it is able to assume that all data that has come from the VMM is trustworthy. Their conclusions are based on tests of the performance of their method against standard access control models. This work is supported by that of Nguessan & Martini (2008) who show that offloading the security processes to some form of 'middleware' reduces the load on the mobile device itself. This increases the performance, particularly with the lower processing capabilities of single mobile devices when compared to that of large servers.

The basic security features within the applications programming environment can also be improved to allow an access control system to be implemented. By default, the Java 2 ME environment does not implement any access control features, and any application with a valid certificate is allowed full access on the system. Ion et al. (2007) implement an

Access Control system for mobile devices by allowing the security architecture to limit an application's scope of control over the device to ensure a more fine grained security policy.

2.4.4 Security Suite

Papers classified as security suite were often concerned with an entire system designed for a specific purpose, such as all the security functionality within a MANET.

Several researchers have looked at creating software solutions for mobile devices that enabled better creation and use of MANETs between them and other mobile devices. Once connected to such networks, mobile devices are exposed to all of the other devices on the network and, by incorporating security features into the network itself, can spread the computational load across the network (Hubaux et al., 2001, Yu et al., 2004).

Trust methods within MANETs are another way to ensure a secure network. Almenárez & Campo (2003) propose a service discovery system for MANETs which is based on a trust network with other devices. In this model, each device keeps a record of the trustworthiness and reliability of each available service and transmits this information on to devices requesting information about the services. This prevents malicious services on the networks from interacting at all with devices since these will always have a very low trust rating.

Huifang et al. (2008) create an authentication scheme for combining 3G services and WLAN networks without compromising authentication or access control principals. This method is based on a public key infrastructure, but using Elliptic Curve Cryptosystems (Koblitz, 1987) which are less computationally extensive and faster than standard encryption processes and so offer better performance for mobile devices.

Bharghavan (1994) created a security mechanism for devices on WLANs which allowed secure communications over a single wireless channel using a shared key between the device and the gateway. This protects the transmitted data from being intercepted by a third party. This method can also be applied to mobile networks rather than just WLANs (Bharghavan & Ramamoorthy, 1995).

Other systems also looked at improving the encryption features or procedures for mobile devices on networks. Kolsi & Virtanen (2004) detail the security enhancements of MIDP 2.0 over the original MIDP (Mobile Information Device Profile), a platform for running java applets on mobile devices. This also enforces the use of HTTPS (Secure HTTP) over HTTP as well as improving the application signing and verification procedure to use X.509 Public Key Infrastructure (PKI) (Adams et al., 2005) certificates for better security.

2.4.5 Availability

The work covering the Availability element is concerned with maintaining the availability of the wireless or mobile network to which the devices are connected. Any network can be attacked using a Denial of Service (DoS) style of attack, and many of the papers looked at protecting MANETs both against this and against malicious nodes on the network either intercepting network traffic or intentionally broadcasting bad data.

Kong et al. (2005) look into MANET availability issues by creating a MANET whereby each node on the network is capable of monitoring the other nodes 'close' to it. If they sense one of these nodes is behaving maliciously they can route traffic around that node preventing it from causing harm. Similarly, Palomar et al. introduce trust levels for

different nodes on the network, with rewards of greater access to network resources of nodes which are shown to be trustworthy members of the network (Palomar et al., 2007).

Malicious nodes can also impersonate legitimate nodes within a network to gain access to resources, or perform DoS style attacks. As such, different methods of authentication nodes often need to be introduced to enable the network as a whole to distinguish legitimate from illegitimate nodes (Mufti & Khanum, 2004, Lin et al., 2007).

2.4.6 Non-Repudiation

Only four of the papers included in the study looked specifically at the issue of non-repudiation on a system, most likely because it is so closely related to the identity establishment or authentication of users. Also there is not a significant amount of research that can be done into such systems, as commonly they involve simply logging the login requests and paths of users within a system.

2.4.7 Others

Works categorised under the Other category were mostly concerned with listing or discussion of security issues relating to a particular technology or concept. For example, Jeong et al. (2008) list generic security requirements for any mobile device, and Carvalho (2008) similarly lists and discusses security requirements for MANET. Other papers looked into detecting malicious activity within a pre-existing network. Both Bae et al. (2007) and Savola & Holappa (2005) present a method to detect a serious of repeated security alerts within a network to show whether certain nodes or peers are continually producing security issues indicating a deeper issue.

2.5 RQ3 Results

The papers found from the Mapping Study were broken down into the types of study they were classified as using the scheme as defined in Section 2.2.1 (Tichy et al., 1995). The majority of the papers published are theory orientated, most often proposing a new model without an implementation. The distribution of study types is shown in Table 4. The sum of the number of publications in this table does not add up to 214 since some of the papers found covered more than one type of study, for example the design of a system that also included an empirical study.

Type of study	Number of Publications
Theory based	123
Design and modelling	75
Empirical	42
Other	13
Hypothesis testing	0

Table 4: Systematic Mapping Study results broken down by study type

Furthermore, in order to research the sort of empirical data relating to mobile security, a further investigation of the papers categorised as Empirical was carried out using the definitions by Budgen (2008). The results of this are shown in Table 5.

Type of Empirical study	Number of Publications
Benchmarking	37
Quasi lab experiment	4

Case example	2
Case study	1

Table 5: Types of Empirical study found from the Systematic Mapping Study

Of the empirical studies found, the majority were performing benchmarking tests related to the performance or speed of the model in question. This is the perhaps the least complex type of experiment to perform, with 37 of the 42 empirical studies using it, since it can be done in a laboratory setting, and requires little outside participation, which may explain its prevalence in the results.

Since the majority of the papers performing empirical studies look into the Benchmarking aspect, there is less likely to be work into the usability aspect of the methods. Benchmarking can be used to show how much better one or multiple methods can perform when compared against a baseline, however, they will offer little insight into the level of usability associated with any of the methods, since they will be primarily concerned with the speed of the method, or how secure the password is or can be.

2.6 Threats to validity

The main threat to validity of the Mapping Study is the limitation of 150 papers from each of the digital libraries. This was done due to the practicality of dealing with the thousands of results which were returned from each library. However, it means that there are likely to be a substantial number of relevant papers missed from the study, particularly relating to authentication.

When starting the Mapping Study, the focus of the research was more in the area of Access Control, rather than Authentication, and so the validation searches were in this area as well. Due to this, an additional literature review, although not systematic, had to be performed into the area of authentication, and this is detailed in Chapter Three. This is also the main reason why some of the important papers included in the latter stages of this thesis were not found through the Mapping Study, but through the additional literature searches performed afterwards.

Another threat to the validity of the study was that only a selection of the available digital libraries was used. Papers published on Science Direct or Web of Science were not included in the study, again for reasons of practicality. Many of these papers would have been indexed by CiteSeerX and so the most relevant ones ought to have been filtered in through searching that library, however there is the possibility some were missed altogether.

There was also a significant gap between finishing the study and the further work which resulted from it. This gap meant that papers which would otherwise have been included were not, and so additional searches would be required to locate these pieces of work. As such it was decided that any further work in this area would require a new literature review be performed, and this is detailed in Chapter Three.

Many of the papers included in Chapter Three, and used throughout the rest of the thesis, were not found during the Mapping Study. This is due to a combination of the reasons already stated above, and the fact that the Mapping Study was performed at a very high level, security on mobile devices as a whole, compared with the more specialised literature

review into authentication methods on mobile devices. This effect of missing some papers over the course of a Mapping Study due to the high level nature of the search string was also noted by Kitchenham et al. (2012). Due to the breadth and size of the overall field, it is perhaps not surprising that some of the papers used later in the thesis were not included in the Mapping Study.

There are also issues with the titles of some of the papers. For example, the original paper used describing the DrawASecret method (Jermyn et al., 1999) is entitled 'The design and analysis of graphical passwords'. Based purely on the title alone, this would not have been included in the Mapping Study since none of the words in the title were included in the search string. Although many digital libraries will also search using the abstract and keywords for the paper, the lack of keywords in the title is likely to decrease the calculated relevance of the result possibly pushing it below the cut off limit on the total number of papers included. This is a further reason why it was felt an additional literature review was needed into the area of authentication on mobile devices.

2.7 Identifying Gaps

During the course of the study, it became clear that very few of the papers being published included any kind of usability assessment of the methods or systems being proposed. A more in-depth analysis was therefore carried out on the papers found from the study to determine how many of them performed a usability analysis, or considered the usability of the final system in their proposal.

Only eight papers mention usability of the security systems investigated. Efficiency of a security system is one of the major factors in the usability of the system. A faster login procedure for instance would obviously benefit and be more widely used, and similarly a security system that puts less of a strain on the mobile processor is less likely to intrude upon the user's experience of the device. For example, Liang & Wang (2004) looked into this but without a specific emphasis on the usability of the application as a whole. Similarly, the idea of combining authentication across multiple mobile networks is principally a usability issue, although the papers found looking into this idea were concerned purely with the security aspects of such an implementation (Wang et al., 2002, Gross, 2006, Ravishankar & Harishankar, 2008).

One aspect that was almost completely overlooked relates to the ability of a user to interact effectively with a security system. This could be considered to be just as important as the security of the system itself, and only Buennemeyer et al. (2008) considered their system in this context. The weak links of any security setup are the users themselves, who are likely to simply switch off security features they find too complex or too time consuming to properly understand and use (Balfanz et al., 2004).

Security features such as single sign on would, by their very nature, make a mobile device easier to use and some papers do mention in their proposed system or evaluation that single sign on would be more usable than standard models (Al-muhtadi et al., 2002, Al-Qayedi et al., 2004, Chen & Sivakumar, 2005, Hashemi & Soroush, 2006, van Thanh et al., 2008). However, only three papers directly considered the usability of the systems they had designed or implemented, (Buennemeyer et al., 2008, Sun et al., 2008, Decker et al., 2004). Both the papers by Decker et al. and Sun et al. relate to the use of an NFC token or device

to allow automatic sign on to a mobile device simply by the user being within range. In the other paper, Buennemeyer et al. (2008), a system is developed to detect intrusions onto a user's mobile device, and a usability study is performed into how much the system affects the usability of the device, both in terms of the additional processing load, and how easy the system is for the users to understand.

2.8 Conclusions of the Mapping Study

The Mapping Study highlights the lack of research in the area of usability of security mechanisms on mobile devices, both in terms of theoretical consideration and empirical studies into the usability of the systems. It also shows that most of the research has been done in the area of identity establishment and authentication.

The Mapping Study also highlights that the majority of papers looking into identity establishment use the term authentication rather than identity establishment when referring to login methods. There are however other works which define access control as the use of credentials such as a password system (Coulouris et al., 2012: 30, Elmasri & Navathe, 2011: 837) and there is some confusion as to which is the correct usage of the term. Since systems such as these, requiring some form of credentials from a user, are generally referred to as authentication methods, these types of method are referred to as authentication methods for the rest of this thesis.

The lack of research into usability is confirmed in the types of empirical studies performed. The majority of the papers which did include some kind of empirical study reported a benchmarking test on a new, or alternative, method. Whilst these studies show how one

method may be more efficient or effective than another, they do not be able to give any indication of the level of user satisfaction the methods provide. Between this and the general lack of work investigating usability as a concept when relating to mobile device security, there is scope for more work into this area.

The limitations of the mapping study suggest that a further, more detailed and focused review of the literature into identity establishment methods or approaches is needed to provide sufficient background for the research described in this thesis. This further literature review is detailed in Chapter Three.

.

3 Chapter Three - Authentication on Mobile Devices

This chapter provides a more detailed review of the literature on authentication methods for mobile devices, since this was highlighted as an area where research was being done in the Mapping Study described in Chapter Two. From the methods discussed here, the three graphical authentication methods, Awase-E, DrawASecret and PassPoints, are chosen and used in the experiments discussed in more detail in Chapter Five.

Authentication methods can be divided broadly into three categories depending on whether the information being used to authenticate users is something they know, something they have or something they are (see Section 2.4.1). However, for mobile devices in particular, the methods available are limited by the characteristics of the hardware and the capacity to attach peripherals such as fingerprint scanners.

Something a person knows is the least affected method in terms of implementation on a mobile device. The only limitation produced by a mobile device on this is the method by which the user inputs the data. Some mobile devices such as smartphones will often not have a physical keypad at all, instead only having a virtual touchscreen keypad, and non smartphones will only have a numeric keypad with the numbers doubling up as letters. This can mean that it is particularly difficult to input alphanumeric passwords accurately onto mobile devices. Some mobile operating systems, such as android, allow for password entry but display the last character input to allow the user to confirm they have pressed the correct key. This introduces a significant security risk should anybody be watching the screen, although could be said to be a necessity given the nature of the virtual keypad the password may have to be entered on.

The other two methods of authentication are severely limited by the hardware available to mobile devices. A desktop or laptop computer can have extra devices connected to it by USB or other means to allow for authentication (for instance a fingerprint reader), however a phone would require this functionality built in to it.

Something a person is relies on biometric data about the user in some way. Mobile phones will normally only have one device capable of visual input, the camera. Thus the ability of the device to use this to obtain biometric data will depend strongly on this camera, and so the application would only be available on models which met the hardware requirements. However, several methods now exist to allow a laptop to use the webcam to authenticate the user through a facial scan, and the new version of Google Android's smartphone operating system has a similar method for authentication available. It has also been shown to be possible to use a handwriting tool to distinguish between users (Humm et al., 2009), or even through the gait of the user as they walk to authenticate the user (Derawi et al., 2010).

Technologies such as AugmentedID (TAT, 2011) allow a user to take a photograph of a person which can be used by the device to automatically discern who the person is from its database. There are already methods such as the FaceR Credential Me (Animetrics, 2011) application which can recognise a user when they are using a phone and allow or deny access to secure sites based on whether the application recognises the legitimate owner of the phone.

Something a person has are often token based systems which use an RFID or NFC card in the device itself to perform an action or authorise a transaction to be made. These are

increasingly being used for mobile payments in shops to simplify the process of paying for small items.

Token based systems can also rely on Bluetooth enabled tokens, or to some extend the range of RFID tags. Some of the methods found from the Mapping Study could apply to mobile devices as well, but may require additional hardware being attached to the phone so Bluetooth would be the best suited if it is to rely on existing phone models and features. The Bluetooth connection would have to be switched on for longer periods of time though, which could represent a significant drain on the battery life over any period of time. There have also been documented a number of security flaws within Bluetooth, such as through spoofing the unique ID of another Bluetooth device, or through a man in the middle style of attack to obtain the encryption key of a session between two devices when it is sent unencrypted at the beginning of the session (Hager & Midkiff, 2003, Shinder, 2005). As such, it is felt this is not a viable alternative given the currently available technology.

3.1 Graphical Authentication Methods

Whilst there are a lot of papers publishing different methods of authentication for mobile devices, only a few are actually in use within the mobile community. These cover the different types of authentication (biometrics, tokens etc) and show that the fact the devices are small, often with limited hardware, does not mean that the standard models or methods of authentication cannot be applied. There is however very little research into how many people regularly use these security features, either in personal or business environments.

All phones come with at least a basic PIN unlock feature which can be used to either lock the handset or access to the SIM card. This involves the user setting a 4 digit pin and the phone can only be unlocked using this PIN. A PIN has been the main method of accessing money through cash machines since their invention and there have been various attempts at looking for alternatives which are either more secure or more memorable (Moncur & Leplâtre, 2007). PINs as a way of protecting bank accounts are relatively insecure by themselves, although features such as chip and pin and the blocking of cards after three incorrect attempts avoid some of the main issues with card fraud.

A PIN itself is only four digits long, and research has shown that many people still use relatively easy to guess combinations (such as 1234 or based on easy to access personal information such as the year of birth) despite how important a resource it protects. For example, if a bank card is stolen along with something detailing a person's birthday, then a thief would be able to get access to the bank account for between every 11 and 18 wallets stolen if they based their PIN guesses around this date (Bonneau et al., 2012). A modern smartphone allows users to download and install a wide variety of applications onto their phone in addition to the programs that come with it on purchase. Although obviously unable to operate on phones without the right technical specification, there are more graphical methods of authentication available that do not need extra hardware which may prove easier for users to use. Biometric authentication methods would need some form of input, such as the camera for facial scans, or a

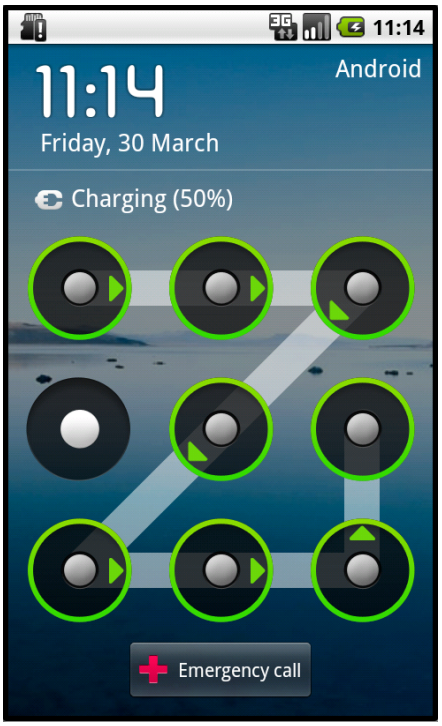


Figure 1: An example Pattern Unlock screen for Android Devices

fingerprint scanner for fingerprints.

New smartphones have begun to offer alternatives to a standard PIN. Both android and iPhones allow a user to choose a proper alphanumeric password to lock the device. Furthermore, android phones are additionally capable of using a pattern unlock method. The unlock pattern involves a grid of 9 points in a 3x3 matrix structure which can be joined up to form the required pattern, an example of which is shown in Figure 1.

A large set of people however will have no security on their phone at all, with one recent survey showing more than half of people surveyed did not use any security on their phone at all (Confident Technologies, 2011). This is a change from an older survey which showed that 82% of people surveyed used the SIM card level Pin lock, although only 15% used the phone security code (Kowalski & Goldstein, 2006). Participants on the more recent survey were exclusively using tablets or smartphones indicating that although these devices are capable of more security functions, people's attitudes to security on them are worse. The most cited reasons are the inconvenience of having such a system between them and using the phone, or the belief that such a system is unnecessary for them (not storing any important details on the phone).

Another recent experiment showed that even when an unsecured 'lost' phone was returned to the owner, the data on it had been accessed in some way in 96% of cases (Symantec Corporation, 2012). Even if this is merely down to human curiosity, any passwords, for example, which are stored on the device could then be in the hands of another party.

As phones become more complex and able to perform far more tasks than older phones, such as personal banking or shopping, the need for higher levels of security on the device will increase. Having said that, there will always be a subset of users who do not wish, or are unable to use the more advanced features of their phones for whom this is not an issue.

Knowledge based authentication methods can be broadly categorised into three separate types, recognition based, recall based and cued recall (Raaijmakers & Shiffrin, 1992, Biddle et al., 2011), which are described in more detail in Chapter Four. A set of typical graphical login methods from each of the three categories, found during the course of the literature review, is described in the following sections.

3.1.1 Recognition Based Methods

All of the methods detailed in this section are recognition based methods which require the user to only recognise their password from a given set of images or letters, rather than recalling it in its entirety such as in the case of a normal password.

3.1.1.1 Awase-E

Figure 2: Example login using the Awase-E method

The Awase-E method (Takada & Koike, 2003) uses a series of images pre-chosen by the user in order to authenticate them. Selecting the images is done through the use of the Gallery method within the Android operating system whereby all of the images available are presented to the user

at once requiring the user to scroll through them to see the full set. A typical screenshot of this method is shown in Figure 2. Awase-E is a purely recognition based method, and as such ought to be easier for people to use than methods that are entirely, or in part, recall based.

To log in, the user is presented with a series of screens, on each of which is a set of images, how many of which is determined by the specific implementation of the method. Of these, at most one of them will be one previously chosen by the user to use as their pass images. It is also possible that all of the images chosen will be decoy images, and so in this situation the user has to choose the 'No Images' option. If the user manages to choose their images successfully out of the decoys on four consecutive screens then they are logged in, otherwise they have to restart from the first screen.

At any point in the process users can to reset and start over from the first screen. If they make an unsuccessful choice of images they are also sent back to the beginning to try again, but are given no indication as to whether or not their chosen images were correct. This helps to prevent a trial and error style attack on the system whereby an attacker could simply try all the images on the screen until they find the correct one and make a note of it. The addition of the 'No Images' option also means that the pass images will not always be displayed on a page, so an attacker would be less likely to be able to see which images are the correct ones by cycling through the login process multiple times, as there is no guarantee the images will always appear.

Awase-E is partially based on the concept from Déjà vu (Dhamija & Perrig, 2000) which uses randomly generated images and patterns as the authentication images rather than

photographs which makes it more difficult for the passwords to be written down somewhere where they could be stolen. It has also been shown more recently that people prefer actual images of objects rather than drawn pictures when authenticating using pictures in this way (Chowdhury & Poet, 2011).

Awase-E has in itself a limited password space. When logging in the user is presented with only 12 of these images and the option to choose no images. This means that through a process of random guessing, an attacker has a $(1/9)^4$ chance of guessing the right password. This is actually marginally worse than a basic 4 digit pin where there is a $(1/10)^4$ chance, and substantially than recall or cued recall based methods.

3.1.1.2 Image Pass

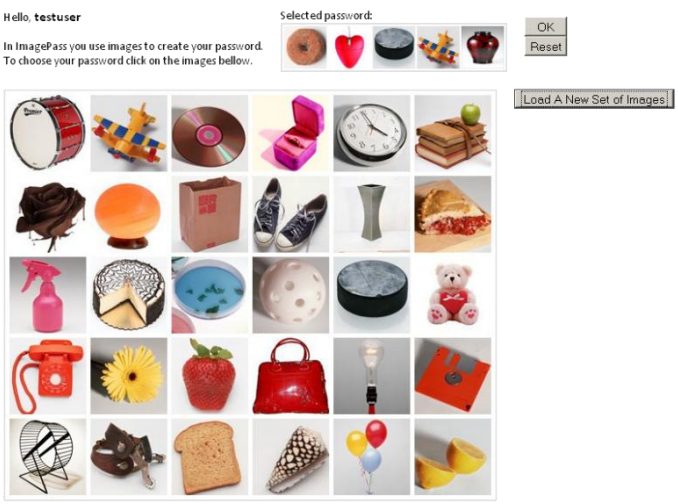


Figure 3: Example login using the ImagePass method

Image Pass (Mihajlov et al., 2011, Mihajlov et al., 2011) is a web based system which uses the recognition of images in a similar manner to Awase-E (see Figure 3), but is designed so as not to have the lessened security associated with having a small set of images. It does this by following the security principals created in Mihajlov et al. (2011), such as preventing the username being used for authentication being displayed, and not allowing the password space to be too small. Their system also uses a two factor authentication method whereby the images being sent are assigned a unique random number at runtime.

The data then transmitted back to the server is therefore unique for each attempt at logging in, so if it is intercepted, it would not reveal any information about the password.

Figure 4: Screenshot of logging using the PassFaces method

3.1.1.3 PassFaces

The basic concept behind PassFaces (Brostoff & Sasse, 2000) is the same as that behind Awase-E. Users are presented with a series of screens containing images, which they are required to pick their own images out of the decoy sets, an example of which is shown in Figure 4. However instead of random images PassFaces uses photographs of real people's faces as the pass images, as there is research to show that humans are better at recognising faces than they are at recognising images, be they randomly generated or photographs of real objects. PassFaces has been implemented commercially and there has been a significant amount of work studying the usability and effectiveness of the system has been performed (Brostoff & Sasse, 2000, Real User Corporation, 2001, Real User Corporation, 2004).

3.1.1.4 Use Your Illusion

Figure 5: Example of how the User Your Illusion method distorts images

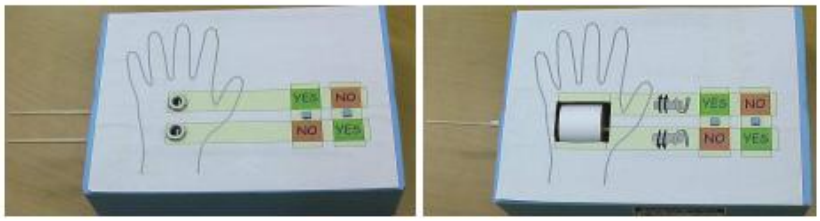
The Use Your Illusion scheme (Hayashi et al., 2008) uses distorted images as the password for authentication. The user first chooses their own relatively high resolution image from a set (or uploads their own), and the software distorts the image to the point where it is unrecognisable, unless the user has seen the original image. An example of this is shown in Figure 5. When logging in, the user is presented with a series of screens with decoy images on them (also distorted from a set of stock images) and the user has to choose their images from amongst them. The scheme has the benefit of not being restricted by the size of the device screen being used, and would work equally well on either a mobile device or desktop, with the authors evaluating the method on a Nokia phone.

They tested their model in relation to how well users of different ages and genders can recognise their images. They then retested users a week after the initial setup and found that when users had selected their own images, they were still able to recognise the distorted images and so log on successfully.

3.1.1.5 Undercover

The Use Your Illusion scheme is extended to use tactile input to produce a system called Undercover (Sasamoto et al., 2008), which is intended to be resistant to observational attacks. Onscreen, the user is presented with a series of images, one of which will be one of their own pass images, and has to choose Yes or No depending on whether they recognise the image. However to choose which of the images is the correct one, rather

than select on screen, the user has to select a button on a box in front of the monitor.



(a) Contact-based prototype (b) Friction-based prototype

Figure 6: Example of how yes and no answers are given using the Undercover method

Which button is associated with Yes and which with No at any given time is determined by some sort of tactile feedback being given to the user's other hand (for example in image (b) of Figure 6, scrolling the wheel up would indicate Yes, whereas scrolling the wheel down would mean No). This would prevent an attacker from seeing which option was being selected for each viewed image.

The authors show how users were able to cope with the two simultaneous sources of information well and logged in with few errors, which would decrease over time as they became more familiar with the system. The time taken for people to log in the systems is shown to be similar to that of other graphical password schemes.

3.1.1.6 Cognitive Authentication Scheme

The Cognitive Authentication Scheme (Weinshall, 2006) is intended to be resistant to both shoulder surfing and spyware style attackers. The user is initially given a large (~100) portfolio of images and is required to memorise them. When logging in, the user is presented with a screen filled with different images, and is required to follow the path of the images from their own portfolio from the top or left side of the screen to the bottom or right. Figure 7 shows a large grid where the user would have to follow the path of their

images across the images and pick the correct number at the end corresponding to the end of their ‘path’

Figure 7: Login method for the Cognitive Authentication Scheme

Once they have reached the end of the path they choose the correct image, and are asked a multiple choice question asking which was the correct path taken. This is repeated several times until the user has performed enough tests for the system to calculate it is sufficiently unlikely the answers were chosen randomly, after which the user has successfully logged in.

3.1.2 Pure Recall Based Methods

The methods detailed in this section are pure recall based methods. Users are required to remember the entire password from scratch without any clues or hints.

3.1.2.1 Blonder's Method

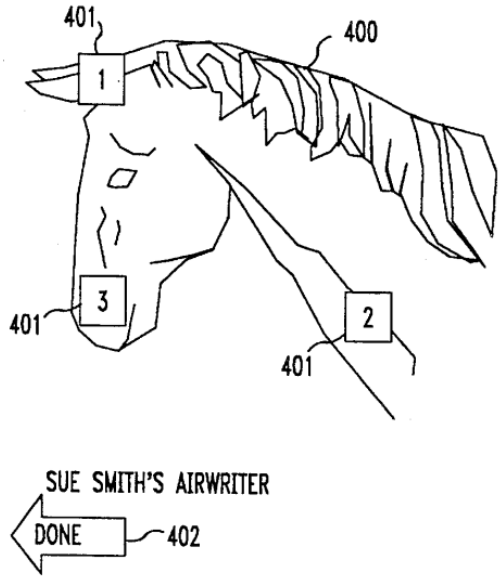


Figure 8: Example login using Blonder's method

Blonder's method (Blonder, G.E., 1996) is one of the first graphical based authentication methods and involves presenting the user with an image of a face or object, and requiring them to tap on certain areas of the screen to act as a password for that user. There would be a certain area of uncertainty around the chosen area of the screen to allow for the user to deviate slightly. The order the areas of the image could also become required here to add an additional level of security to the password. Figure 8 shows the original example of using a horse as the base image, and squares 1, 2 and 3 the areas on the screen the user would have to tap in order to log in.

3.1.2.2 DrawASecret

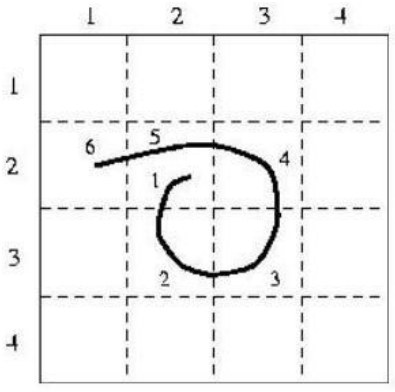


Figure 9: Example DrawASecret password

The DrawASecret, or DAS, method (Jermyn et al., 1999) involves a user drawing a pattern on the screen which acts as a password for that user. A grid structure is displayed onto which the user overlays their own pattern. The grid allows for some deviation in the exact replication of the pattern allowing the user to remember a basic shape rather than the exact drawing. As long as the replication passes through the same grids as the original password then the user would be authenticated. Any combination and any number of lines and points can be used, and

obviously the more complicated the pattern used is, the more secure the stored password will be. An example pattern is shown in Figure 9.

DrawASecret is a purely recall based method since the entire pattern drawn by the user must be remembered by the user. For this reason, this method is expected to be harder for participants to use and remember their password. However, the password space here is technically infinite since there is no limit on how many lines can be drawn on the screen. After a certain point though it is unlikely a user would be able to accurately remember the password well enough to be able to log in using it. This is the case for the other drawing based authentication methods such as PassDoodle and PassShapes.

Circle or curved lines which pass too close to the gridlines or intersection of gridlines can however be difficult to reproduce because only small movements are required to move the line into a different grid and so substantially change the password.

This method is similar to the graphical unlock screen on modern Google Android based phones but there are some significant differences. The Android unlock screen involves joining 9 points in a square arrangement without using the same point twice, and in one continuous motion, without allowing for multiple lines or patterns to be drawn as is the case with DrawASecret.

3.1.2.3 BDAS (Background DrawASecret)

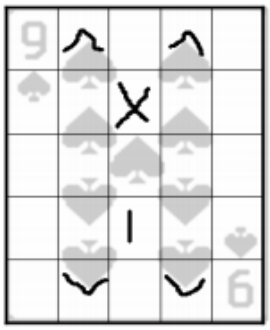


Figure 10: Example of a DrawASecret password using a background image

BDAS (Dunphy & Yan, 2007) is an extension of DAS which uses a background image instead of a basic grid to help users remember the pattern they have drawn, or to guide them into drawing a memorable pattern based on the background image. Figure 10 shows how this could be used to remember the pattern using a playing card as the background. The lines are drawn specifically to match certain features and patterns on the face of the card.

3.1.2.4 PassDoodle

PassDoodle (Goldberg et al., 2002) is again similar to DAS, but uses no grid structure for users to match their drawings up to, just presenting the user with a screen on which to draw a freehand image. Additional features such as colour and thickness of the drawing line were added to improve the security of the finished drawing. When logging in, a matching algorithm was used to compare the user's drawing with their original attempts to see whether the two drawings are similar enough to authenticate the user.

3.1.2.5 PassShapes

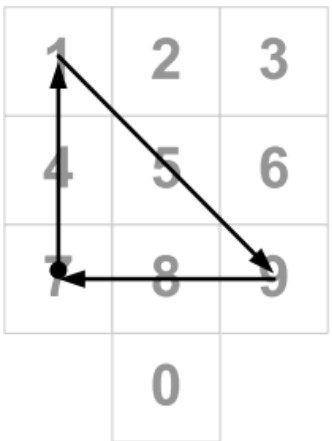


Figure 11: Example PassShape login to remember the PIN 7197

PassShapes (Weiss & De Luca, 2008) converts a user's drawings on the screen to a series of 8 possible directions (north, north east, east etc). This allows each motion on the screen to be easily stored on the device as a series of directions made, so the user has to reproduce these directions in order to

authenticate themselves. This means the user can draw their shape or pattern anywhere on the screen and it will still recognise the correct directions drawn. An example of remembering a PIN using this type of method is shown in Figure 11.

3.1.2.6 PassGo

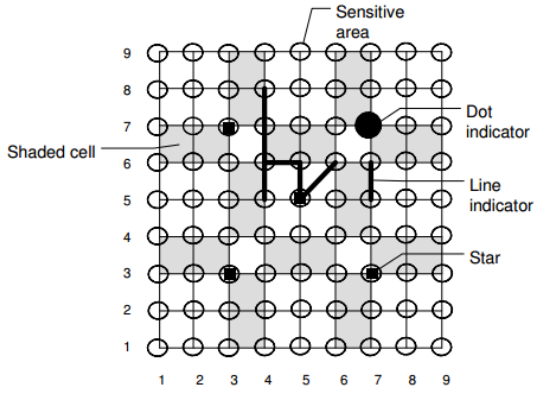


Figure 12: Example of how the Pass-Go method works

PassGo (Toa & Adams, 2008) is intended to remove one of the main usability issues of DrawASecret in that the system can have trouble recognising which grid a drawn line passes through (or is intended to pass through) if it passes too close to the intersection of two of the grid lines. The user is presented with a grid as before, but instead of being a freestyle drawing, the user's movements are 'snapped' onto the intersection of the gridlines to create a geometric pattern from the chosen dots (Figure 12). A less complex version of this has been implemented as the graphical unlock screen on Google Android phones (such as Figure 1) and PatternLock (Tafasa, 2011) for Blackberry phones.

3.1.3 Cued Recall Methods

The following methods are known as cued recall based methods since they rely on recalling information, but certain clues or indications are given to users to help them remember the original information rather than requiring them to reproduce it entirely from memory.

3.1.3.1 PassPoints

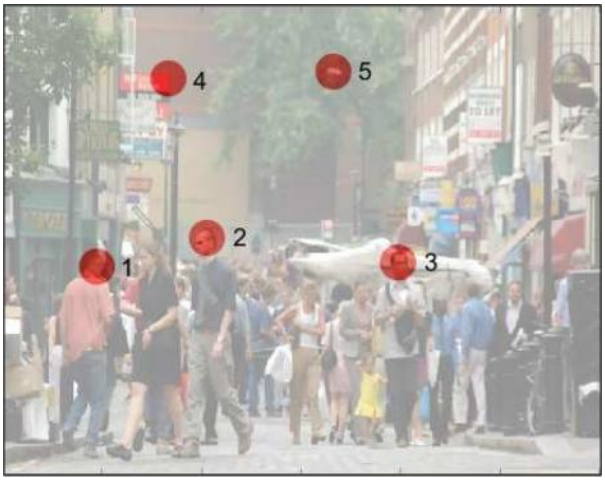


Figure 13: Example set of PassPoints highlighted on a background image

The PassPoints method (Wiedenbeck et al., 2005) is split into two stages, the first requires the user to choose a background image for their password, and the second requires them to choose areas on the screen to act as their password, in a similar manner to Blonder's method. The point the user has chosen is stored in some format on the device, and the user has to then select a point within a certain distance of this (in a circle around it) in order to successfully log in, such as shown in Figure 13.

PassPoints is a cued recall based method since the user is given a cue (the background image) as to where the points in the screen ought to be. From a purely memory perspective, this method ought to be easier than DrawASecret, but more difficult than Awase-E for the participants.

3.1.3.2 Cued Click Points

Cue Click Points (Chiasson et al., 2007) or CCP is similar to PassPoints except that instead of choosing a series of points on the same image, users are required to choose a point on a series of different images. This was later extended to Persuasive Cued Click points (Chiasson et al., 2008) where the system encourages the user to choose a sufficiently

random set of points on the images to stop ‘hotspots’ on images where all the users chose the same point.

3.1.3.3 Triangle Scheme

The triangle scheme developed by (Sobrado & Birget, 2002) which is intended to be resistant to shoulder surfing style attacks. In this type of attack the attacker watches the inputs made by the user whilst they input their password and so can work out the password very easily. Graphical passwords are particularly susceptible to this type of attack particularly when the user has to choose the images off a touchscreen type input since it is easy to see which images the user is choosing, or pattern they are drawing (literally by looking over their shoulder, hence the name).

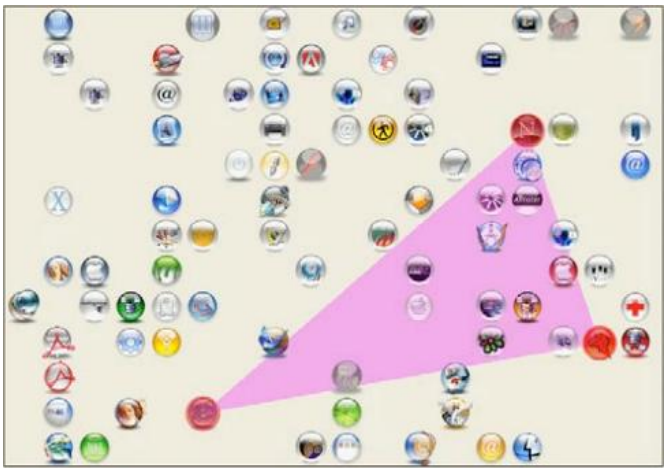


Figure 14: Example of logging in using the Triangle Scheme. Any icons in the pink area would work as one of correct icon for logging in on this screen.

The triangle scheme avoids this by presenting the user with a large number of small icons on the screen. In amongst these icons will be one or more icons the user has previously chosen to be their pass images. The user then has to make a triangle using the other icons on the screen which contain their chosen pass images, such as is shown in Figure 14. Since the user never actively picks their own images, and the triangle will contain any number of decoy icons, it would be much harder for a watching attacker to ascertain which of the icons are the correct ones.

This method is not easily applicable to mobile devices however due to the smaller size of screen. In order to be effective, this method would require a large space from which the user can easily choose the smaller area containing their pass icon. However this would not be possible on a smaller mobile screen without having the images either too small to be visible or having too few images to make an effective decoy set.

3.1.3.4 S3PAS

The S3PAS method (Zhao & Li, 2007) is very similar in application to the triangle method. However instead of using graphical icons S3PAS uses a text based password itself and displays a series of grids of letters and characters on the screen. The user has to find, but not select, three of the characters found in their password from those presented on each of the screens (where the letters and characters are not all presented each time and placed in random positions). These characters make a triangle, and the user is then required to choose a character inside this triangle as their pass character on that particular screen (see Figure 15). This is repeated four times so the user has chosen a 4 character temporary password. This temporary password is then used as their login for the system.

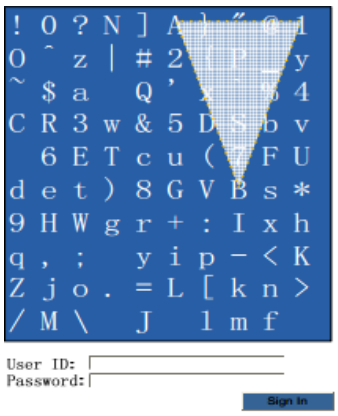


Figure 15: Example login using the S3PAS method for password triangle A1B

Since the user never selects any of the characters from their own password, and the full set of characters from the password will never be displayed on screen together, it would be very difficult for a watching attacker to see what the set of characters used for the basic password would be, and so would be unable to log in using the system were they to try themselves.

This method has advantages over the triangle method on a mobile screen since it is easier to recognise small characters than it is to recognise small images. This would make it more feasible to implement on a mobile device. There would however still be issues due to the fat finger problem which would mean a larger portion of the screen would have to be 'acceptable' as the temporary character for the login, and so make the application more vulnerable to random clicking attacks.

3.1.3.5 WIW

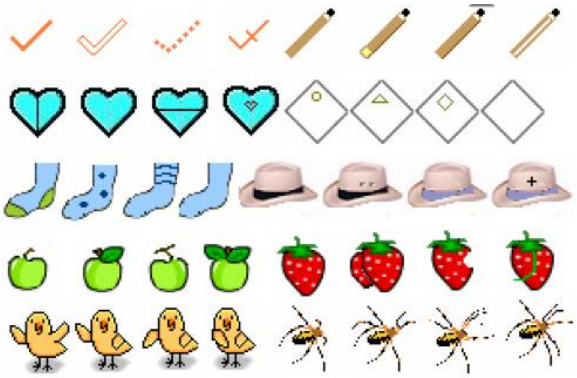


Figure 16: Example of how sets of images are created with minor differences between them

First proposed by Man et al. (2003) and expanded by Hong et al. (2004), the WIW graphical password scheme aims to produce a graphical password scheme which is resistant to spyware and other forms of logging which can be used to harvest passwords. Similarly to Awase-E, the user chooses a set of pass icons which are then displayed to the user on a series of screens along with decoy images. However, each of the images available has slight variations on the image (such as those in Figure 16), so for instance if the user had chosen a kitten as an image, then the colour of the ribbon around its neck might be changed when logging in, with these changes being made randomly. The amount of variation allowed is controlled by the user and the decoy images are also altered on each viewing.

A piece of spyware monitoring the actions of the user would note which image was chosen in order to log in, but that exact image would not necessarily be available next time the user attempted to log in, or if the spyware was automated to attempt itself.

3.1.3.6 Image Authentication

Figure 17: Example login password for the Image Authentication method. The squares highlighted with a black edge are used as the password

Doja & Kumar (2008) propose an image authentication procedure which is intended to be resistant to spyware installed on a machine. In this system the user chooses an image to act as a background for the authentication procedure. Any image, including user uploaded ones, is allowable for authentication. A grid is placed over the image and the user asked to choose a series of grids from the picture. This choice is hashed and stored securely and becomes the password for that user. The order in which the grid squares are chosen in this method has to be preserved. An example of how this would work when logging in is shown in Figure 17.

They evaluated their system with 35 participants to see how well they could remember the passwords over longer periods of time as well as after creating the password in the initial session. In comparison with an alphanumeric password group, the participants using the graphical password were able to remember their password better in the final session four weeks after the original creation.

3.1.3.7 Learn By Use Graphical Password

Herzberg & Margulies (2012) present a graphical password intended to be easy for users to learn to use with no formal training on the system. The system has a number of security features such as increasing the difficulty of the authentication procedure if it recognises

that an attack is in progress. Users choose a set of images to act as their password and are required to choose these from amongst a series of decoy images in a concept similar to that of Awase-E. They also present a web based version of this for any user to try, as well as a fallback mechanism which could potentially be used if the graphical method has failed or is unavailable (Herzberg & Margulies, 2011).

3.1.3.8 Inkblots

Inkblot (Stubblefield & Simon, 2004) involves using images, specifically inkblot or Rorschach style images, as cues for a normal text based login system. Users are presented with a set of the inkblots, such as those in Figure 18, when creating their password and asked to use the first and last letters of the word best describing the inkblot as part of their password. A series of inkblots is presented to the user to produce the full length password.

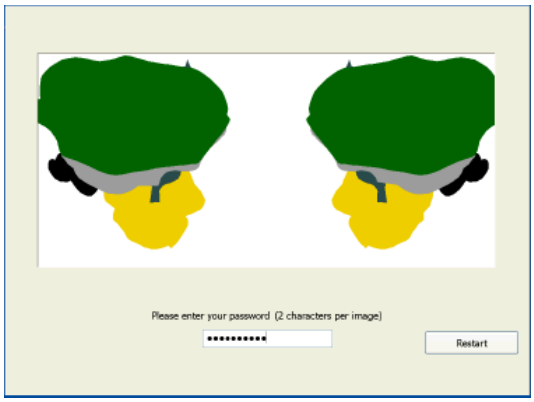


Figure 18: Sample inkblots presented to a user when logging in

3.1.3.9 Jiminy

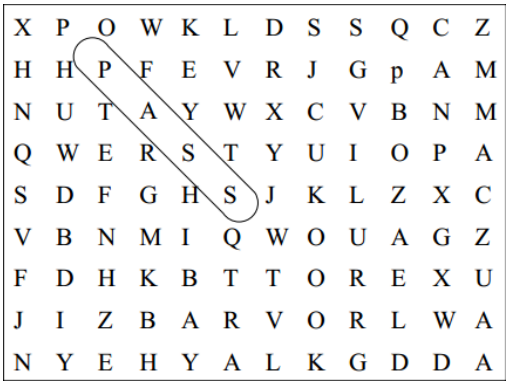


Figure 19: Example of how the password can be arranged in the letter grid using the Jiminy Method

Jiminy (Renaud & Smith, 2001) is intended as a method to help users remember text-based passwords by presenting the user with an alphanumeric grid similar to a word search. The user then has to locate their password within this grid and draw the correct line to indicate where the word starts and begins, an example of which is shown in Figure 19. The pattern of gestures made by the eyes over a grid presented to them represents their password. They performed a study and showed their method to be more secure than PIN authentication (as it is virtually impervious to shoulder surfing style attacks) and that people are able to use it. They did not perform a study on the memorability of such passwords, leaving that for future work into the subject.

3.2 Other Knowledge Based Authentication Methods

There are of course many other methods of authentication which come under the heading of knowledge based methods. The most familiar of these, passwords, has been the subject of a great deal of research over the past few decades most of which has been into the effect of password choices by users on a system, and what steps can be taken to ensure users either choose or are forced into choosing secure passwords.

3.2.1 Passwords

A number of empirical studies have been performed on password security and people's ability to remember them. One study found that 80% of people used only alphabetic passwords and another 80% never changed their passwords, going against all of the guidelines for secure password creation and maintenance. The participants used in this study were computer users within the American Department of Defence and so were required to have passwords for access onto the computer systems there (Zviran & Haga, 1999). Although their work is over 10 years old, it is unlikely that many of the trends found will have changed significantly. Strangely, work from 7 years prior to this shows that only 1 in 5 passwords to be based solely on dictionary words indicating that over time password security might actually be decreasing (Spafford, 1992). This is most likely due to an increasing number of people having multiple accounts requiring a password, and so they choose simpler ones to be able to remember all of them.

For instance, there have been several recent incidents involving the publishing of passwords from different websites. These have shown that a substantial number of people still use easy to guess dictionary words (e.g. 'password', 'link') or simple strings of numbers (e.g. 12345) as passwords (*Programming stuff*, 2007, Calin, 2009, Carey, 2012, Kelly, 2012) . In a corporate environment this type of password behaviour is likely to be prevented by company password policies. This is most likely due to the issue that people have difficulty remembering strong or randomly created passwords. However, it has been shown that passwords created using mnemonics are equally as secure as random passwords, despite being substantially easier to remember security (Yan et al., 2004).

Training users to use random passwords may not be as effective as is often hoped when trying to increase general password security.

(Inglesant & Sasse, 2010) present a more recent study into the use of passwords in a real life scenario. They find that many of the password security policies that have been put in place due to previous work on passwords are too inflexible to allow people to create passwords they can remember, and so negatively impact on the usability of the system and productivity. They point out that with the advent of cloud computing, brute force attacks are likely to become more feasible over time, and so current password policies need to be made more flexible to allow users to choose better, more memorable passwords rather than simply enforcing even more complex random passwords when they are unable to remember the ones given now.

3.2.2 EyePassShapes

In addition to alphanumeric and graphical passwords, there are other more novel schemes to allow authentication. For example, EyePassShapes (De Luca et al., 2009) is a gesture based authentication method intended to be used a public terminals where passwords and other personal details could be stolen if entered directly onto the terminal or device. The method uses the motion of the eyes of the user to create a pattern rather than attempting to work out where exactly the user is looking on the screen.

3.3 Other Authentication Methods

As described in section 2.4.1, there are three general types of authentication based on the type of information required (Wood, 1977). As well as knowledge based authentication, there are also biometric and token based authentication methods, and these are described in the following sections.

Further information and pointers to the literature can be found in a survey of authentication methods which includes a discussion of their advantages and disadvantages (Hamilton et al., 2007).

3.3.1 Biometrics

Biometric methods of authentication are commonly in the form of fingerprint scanners. These have been available on some laptops for several years and allow the user to replace the standard windows login screen with a scan of their fingerprint. However, authentication methods such as facial recognition are also becoming more common as the camera functionality on smartphones improves. For instance, the latest version of android recently released has inbuilt facial recognition software to lock access to the phone itself using the inbuilt camera (Google Inc, 2011).

In a recent review of biometric authentication, it has been shown that simply using a biometric authentication method does not guarantee a secure system, and given the risks involved with storing biometric data from a privacy standpoint, the security on the

database where the information is stored needs to be equally as robust as for any other system (Bhattacharyya et al., 2009).

An alternative type of biometric information which could be used is using a person's gait, or walking style, as a method for authenticating themselves on a device. This would depend strongly on the quality and accuracy of the accelerometers within the specific device (Derawi et al., 2010). The results show that it is possible, and such a method would provide a highly usable unobtrusive method of authentication. However, the then current error rate of 20% is too high to be of practical use.

Even without specific biometric scanners, any device with a touchscreen phone could in theory be used by allowing a user to draw their signature on the screen. Before smartphones had gained a significant amount of coverage or the large user base they do now, work had been done into the feasibility of this concept, with (Narayanaswamy et al., 1999) showing how a whole operating system had been designed on a Nokia device to this end. The signature as a form of authentication is similar in effect to DrawASecret on the assumption that everybody would draw things in a certain way.

3.3.1.1 Fingerprints

(Priya & Rajesh, 2011) present an overview of many of the current uses of fingerprint systems for authentication showing over how wide an area they have been adopted in security systems. They also provide an overview of the literature on this type of security.

Fingerprint scanners for laptops have been available for several years allowing the user to skip the standard login screen on Windows with a scan of their finger. Other research has

shown that users are often happy to use fingerprint scans to log into their computers, and in some cases prefer it over that of the standard username and password approach (Moody, 2004, Heckle et al., 2007).

3.3.1.2 Facial Scans

Facial scans are another method to securely identify a user onto a device without the user having to remember any complex passwords, as long as the device itself has a camera of some form. Yung-Wei Kao et al. (2008) present a method for authenticating a user through face scans from the camera on a mobile phone or device. Their system uses a two stage authentication process, requiring both a password and the facial scan to confirm it is the correct person (and to prevent a photo from being shown to allow access without the owner present). Furthermore, as the password is input first into the system, the system can ignore a user whose password does not match. This would significantly reduce the possibility of a false positive match on the facial scan.

False positives will however always be an issue with biometric authentication since if the system decides the information is provided is incorrect, then there will be no way of providing alternative biometric data. In such cases an alternative login method would be required. Sirlantzis et al. (2010) investigate a series of ways of authenticating users in a ubiquitous computing environment, such as using iris scans, gait analysis and facial analysis. These methods are intended to be an unobtrusive authentication mechanism which would fit better in with the idea of ubiquitous computing than users continually having to authenticate themselves to use the devices around them. All three of the authentication methods can be used at once to provide a stronger authentication then any one of them individually.

Since many of the cameras available on smartphones, or webcams on laptops, are not high performance devices, the chance of the camera not being able to create an accurate enough picture of the person's face is higher than if proper cameras were used. Linlin Shen et al. (2010) investigate this issue as well as looking at how well the microphone would be able to recognise speed patterns as an alternative form of authentication. They show that individually the error rates for the two types of authentication are too high, but when the two types of biometric information are combined this is reduced to 21% which shows it could be useful in the future when the technology improves, particularly with additional algorithmic analysis of the speed patterns.

As with fingerprint scanners, technology has been available on new laptops and smartphones to allow login to be done via a face scan via the webcam. In such cases, the device can allow a correct facial scan instead of the standard password login screen, although manually typing the password is still available as an option should the user not wish to use the feature.

3.3.1.3 Keyboard Analysis

Keyboard analysis of the way people type can be used to authenticate users, since the speed and pattern people type with has been shown to be unique, particularly over large pieces of text. Killourhy & Maxion (2008) look into some of the issues associated with using this as an authentication method, particularly in relation to low resolution timing mechanisms. Since the difference between two people's timings of their typing patterns may be of the order of 10-15ms, low resolution clocks may not be able to distinguish between two people in this situation. By comparing a high resolution clock to some lower resolution ones, they found that many of the lower resolution clocks are not capable of

producing an accurate enough set of timings for a person's typing and so would not be able to effectively be used as authentication methods.

Humm et al. (2009) extend this concept to use a combination of both written word and speech to authenticate themselves. The user is prompted to write some text, be it their signature or a random piece of text, and speak the words they are writing at the same time. The system can then match both the handwriting style and speech patterns to properly authenticate the user. Each of the inputs is analysed independently and only if both the written and spoken methods match will the user be authenticated. This reduces the time taken to log in since both sets of data can be recorded at the same time.

Similar to keyboard analysis is the concept of handwriting analysis where an application is able to distinguish between the handwriting of different people and use this as a method of authentication. This type of biometric authentication is possible on touchscreen devices where the user can draw their own signature directly onto the screen, and there are various applications, such as BioWallet (Mobbeel), which already offer this functionality.

3.3.2 Token Based

Location based authentication involves a user wearing or having some form of NFC device which, when they come into close enough proximity to a terminal they wish to use, allows them access onto the system without typing in any form of password. This has significant benefits from a usability point of view, but obviously is open to attack if the device (even if woven into an item of clothing) is lost or stolen by an attacker.

3.3.2.1 NFC

NFC authentication methods involve a user having the token on them, either in a special card or in some other form where the user would have it close by at all times (various proposals have presented ideas where the NFC chip was included with a smartphone). An example of how this could would be a user (in this case a doctor or nurse) wanting to use a general computer within a hospital environment. Once the user is within range of the terminal or computer they wish to use, the NFC card is detected and the user is automatically logged onto the machine. Similarly, once the user leaves the machine, the NFC chip would no longer be close enough to the computer and so the machine would log out. This approach, although it would improve the usability and ease of access to records, has severe security implications if the NFC device is lost at any point, and the data to which it allows the user access is sensitive information, such as in the case of medical records.

A practical example of using a phone as the NFC device has been demonstrated by Saxena et al. (2011), who look into using a phone as an authenticator for an RFID tokens to stop them being used anywhere, possibly without the user's permission. The user has to touch their phone to the RFID tag before the data on the tag can be accessed.

3.3.2.2 Pass Cards

In a medical setting, the traditional office based password authentication mechanisms can interfere with quick access to often vital medical information. One way to avoid this would be to use a location based authentication system whereby medical professionals can log in seamlessly through active devices in pens on their clothing when they are in range of specific machines (Bardram, 2005). They also point out that the security of a system

where the authentication is too complicated will be nearly completely countered by people circumventing the authentication process.

3.3.3 Comparisons

Braz & Robert (2006) gives an overview of the usability of many different authentication methods, including all three of the main types of authentication: biometrics, token based and knowledge based. They conclude that there is still not enough research into the usability issues associated with all of the authentication methods, and propose that there needs to be a set of guidelines which system designers and researchers can used in order to make new systems more usable, though they do not propose any guidelines of their own to address this.

There is a significant difference between the security required on a desktop compared to that of a mobile device, and the security measures required for each of them. Botha et al. (2009) investigates this in terms of what they offer protection wise from the point of view of a user who is familiar with desktop security procedures. They support the belief that there is a lack of security procedures on mobile devices and that alternative methods of authentication may be required beyond passwords since these are not easy to input on the device. Many of the security issues affecting mobiles do not yet have solutions at all and users need to be aware of the reduced security around file protection and online transactions.

3.4 Guidelines

Many guidelines have been published on how to best keep a system secure. Most of these do not relate directly to mobile devices or phones but many of the concepts presented are just as relevant to mobile devices as the intended systems, be they web or PCs.

Yee's guidelines are aimed at the security side of a system and how it should respond to different actions by the user (Yee, 2002). The security aspects of the system are split into 10 basic areas which are then expanded on to show how each should be implemented. These then cover both the back end of the system and how it should deal with security sensitive tasks thereon and how this information should be conveyed to the user.

Ahituv et al. (1987) present a set of guidelines specifically relating to the authentication of users onto a system. These cover how the data relating to a user should be stored on the device, and the ways in which a user should be authenticated successfully.

Several sets of guidelines exist to allow businesses to model how secure their systems are. Lowe (1997) produced some models for different levels of authentication on a system and how they can be classified. They describe with a model to test different levels of authentication and how well they can stand up against different types of attacks, concentrating on injection attacks where the authentication of one user can have the unintended consequence of allowing a different user access onto the system. Similarly, (Gritzalis & Katsikas, 1996) discuss way of modelling the authentication process to highlight any issues which could be exploited by attackers before the system is built so they can be addressed from the outset.

The work by Burr et al. (2006) is intended to provide organisations implementing electronic authentication methods with the necessary framework and guidelines to implement the correct level of security. It defines four levels of authentication with level 1 being no authentication up to level 4 requiring strong cryptographic keys to ensure the user is who they say they are. For the most part, the authentication procedures defined in this thesis are level 2 only requiring knowledge of a password (be it graphical or otherwise) for the user to gain entry.

How the data is stored on the device is another important factor in authentication. Should the device be stolen, or the database compromised, if the passwords are stored in plain text then the attacker would have instant access to all of the users on the system as a whole. Abadi & Needham (1996) discuss some of the uses of encryption and how and when they ought to be used. They present several principles for the use of encryption on systems, concentrating on why and where it should be used rather than simply on ensuring everything is strongly encrypted. They also discuss how encryption procedures, namely one way functions like hashing, can be used to generate random numbers as used in the implementation of the authentication methods used in Experiment 1 described in Chapter Five.

The next chapter describes some of the research on usability looking at both general literature and that relating more specifically to mobile devices. Research on usability, along with the work presented in this chapter, form the basis for the development of the criteria used to evaluate the three chosen methods in the experiments detailed in chapters five to nine.

4 Chapter Four - Usability

This chapter reviews usability as a whole, covering the main definitions of usability and how it can be assessed. This includes not only the published work, but also the international standards for usability according to the ISO (International Organization for Standardization). It begins by looking into usability in general along with the psychological factors which will affect how a system as a whole is perceived and used. Also presented are examples of how usability concepts have been applied to the design of actual applications, both for business and non business environments. How the usability of mobile devices and applications can be assessed is then discussed, specifically relating to the usability challenges of the small screen and linking in with the work done in the previous chapter on mobile security as a whole. Finally, some currently available methods for evaluating the usability of applications, both mobile and non mobile, are discussed.

In one of the seminal papers on modern usability, Bevan (1995) discusses what it means for a system to be usable, contending that a system is usable if the quality of usage of the systems is high. Historically, quality of software systems is concerned with how well it meets the original intention. However it is suggested that this should also include how well the system can be used to meet its goals, whatever they may be. Bevan also suggests some ways of measuring the usability of a product along with suggesting that user-centered design be incorporated into a project the whole way through rather than just at the end, so that any major issues with the system can be addressed early on and avoided. Similarly, the work by Norman (1988) is considered to be another of the seminal works in modern usability design, and relates mostly to how users actually approach using a system, from a psychological standpoint rather than a technical one. Norman strongly advocates the use

of user-centered design in the design process of any system in order to find potential issues before the system is released to its intended market.

One way to measure the usability of a system in a systematic way is the MUSiC method (Bevan & Macleod, 1994, Bevan, 1995). MUSiC is intended to provide a standard way of measuring the usability of a system, and provide tools to help resolve issues. It is shown how it can be used to make and meet goals for usability within a system and make an effective measurement of the system's usability from more than one standpoint, such as novice and expert users.

A different way to look at the usability of a device is to split it into three separate areas: user interface (buttons, display, audio etc); external interface (support, accessories, other software); and service interface (availability, interoperability) (Mika & Röykkee, 2001). The researchers here note that previously they had found that many users want interoperability between devices as well as a constant and reliable connection to the mobile/wireless network as a whole. Similar to comments made by Oppermann (2002), they note that devices should also not stigmatise users such as elderly people as "disabled" as this is off-putting to them. Rather the device should be made generally more usable, and this is likely to help the majority of people. Much of their work is designed around this premise and shows how an interface should help users, as well as cater to different levels of user (novice, casual, expert etc), and how user centred design should be included from the beginning of the design process to allow the users themselves to help show what most needs doing on the device before it is released.

4.1 Definitions

Since the term was first coined, usability has been described in many different ways, often related to ways of measuring it. The basic idea behind this is how well a system allows users, both novice and expert, to use it easily and efficiently.

Usability has been defined as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO/IEC, 1998). This standard is used as the basis for usability studies across the world by splitting the area of usability into three separate categories.

1) Effectiveness: Measures of effectiveness relate the goals or sub goals of the use to the accuracy and completeness by which these goals can be achieved.

2) Efficiency: Measures of efficiency relate the level of effectiveness achieved to the expenditure of resources. Relevant resources can include mental or physical effort, time, materials or financial cost.

3) Similarly, if any application is too demanding for a user (for instance requires remembering too much information in order to log on) this would lead to low efficiency of the system, both in terms of the mental effort required to log in and the extra time required to perform a more complicated procedure.

Satisfaction: Satisfaction measures the extent to which users are free from discomfort, and their attitudes towards the product.

These basic definitions of the three aspects of usability are used in both of the ISO standards for usability, 9421-11 and 13407. However, these two standards approach the issue of usability from different standpoints. ISO 13407 (ISO/IEC, 1999) considers the design aspects of usability, specifically in the form of prescribing a method for a rigorous Human-Centred Design. If followed correctly, it would imply that a system ought to be fully usable, since the users will have effectively been involved throughout the entire design process. ISO 9421-11 looks at usability from the standpoint of analysis of an already designed system and measuring its usability using whatever scale is appropriate to the system.

Work has shown that the two guidelines are not completely complementary to each other (Jokela et al., 2003). Specifically they show that through use of the Human-Centred Design in ISO 13407, the final system will not comply fully with the usability specifications set out in ISO 9421-11.

From the literature review relating to mobile security, there was found to be a lack of significant research into the area of the usability associated with the security systems. For instance, from just the Mapping Study performed, only eight papers were found to consider the usability of the system at all, with only three papers doing any form of evaluation or study into their systems.

Efficiency

In the case of an authentication/login system, this would be defined as the ability for the user to successfully log in without any cases of the wrong password being input and the system allowing the user on, or the user input the correct password but being denied entry.

Effectiveness

In terms of a mobile device, the efficiency of the system is reliant on the performance and speed of the device itself. Due to the relatively limited amount of resources, computationally heavy applications will lead to long delays in using any application and so would make many users impatient. Research has found that one of the main reasons users currently do not use the security systems already in place on mobile devices is the extra 'hassle' of needing to perform tasks before being able to access resources (Balfanz et al., 2004).

User Satisfaction

User satisfaction is almost entirely a subjective experience based on each user, and so often requires a usability test to measure adequately. Users can reflect on their experience not only of using the system in question, but also on the other aspects of usability (efficiency and effectiveness) as they experienced them. Objective measurements can also be taken of the user whilst using the system (e.g. negative comments or frustration) and a measure of the satisfaction can be inferred from the number of errors made by users. For instance, if a system has a high level of errors then it is likely the satisfaction will be low as users are unable to use it effectively.

Some studies have shown that when a user is presented with a security system they are unable to cope with, or simply do not like, they are inclined to turn the security system off, not use, or invalidate the security by using very simple passwords (Adams & Sasse, 1999, Balfanz et al., 2004, Karlog et al., 2004, Dingledine & Mathewson, 2006). This work was done in relation to passwords on desktop or laptop systems but the thought process could equally well apply to mobile devices. Once the user has decided they do not want to use any form of security they are opening themselves up to attack unnecessarily.

4.2 Usability Guidelines

In addition to guidelines and criteria for evaluating usability on mobile devices, several sets of guidelines have been produced in order for developers to measure their applications or devices. Many are based on the same original principles such as the original ISO usability specification (ISO/IEC, 1998) and the Usability.gov guidelines for websites (HHS Web Communications Division, 2009), which are themselves based on a wide range of usability standards and practices.

Some of the actual security issues facing modern computers are discussed by Yee (2004). Issues such as users' mental models of security and how much they understand (or do not understand) security concepts are investigated along with how this can lead them to make poor decisions. When and where these decisions are made are then matched up to aspects of operating systems such as Windows, Linux and MacOS and how they can be addressed in these situations. The guidelines created by Yee (2002) were taken as one of the main sources for the criteria developed to test the usability of mobile applications in the experiment detailed in Chapter Five.

The guidelines for usability on handheld or mobile devices can also be adapted from ones designed for devices with larger screens. Weiss (2002) compiled a set of guidelines specifically for this purpose, along with suggesting some different methods for usability evaluating an application using these guidelines. These include having an audio/visual record of the tests performed and using logging software to record users' actions.

Combining usability and security without losing the benefits of either is the goal of many researchers. Johnston et al. (2003) propose a set of criteria/guidelines for secure HCI interfaces and show how the (then current) Windows XP firewall did not meet many of them. From this they propose several changes to allow normal users to be able to benefit from increased security and understanding of what the system is meant to do.

It has been shown that usability testing is better done out of a lab environment if it is to most effectively identify usability issues with a device or application (Gorlenko & Merrick, 2003, Duh et al., 2006). In addition to this though, it is also important to make sure the system is evaluated by users with different ability levels, to make sure that expert users are not annoyed by too simplistic an interface, and vice versa (Oppermann, 2002). Oppermann also presents a generic set of guidelines intended to cover aspects of both usability and security within a system.

It has also been suggested that the original three elements of usability, efficiency, effectiveness and satisfaction, would be better looked at together as usability as a whole rather than broken down into their constituent parts, particularly when testing (Frøkjær et al., 2000). Only a weak link is found between the three aspects and so suggest that where possible they should be considered as entirely independent variables, and so studies which only look into one or two of them cannot be considered to be doing a full usability study since they will have missed at least one aspect of usability. They show that many of the usability results from recent CHI papers do this however and a substantial number of papers have results which will be incomplete as a result.

Conversely, in a much earlier study, a direct relation between the ability of users to perform a task and their satisfaction about using the system was found, although it is noted that there are examples where this is not always the case (Nielsen & Levy, 1994). Intuitively this makes more sense since if people are able to use a system they will be more inclined to like it and vice versa. As such, a user's opinions of a system should be a vital part of any usability testing and so should always be gathered.

The Usability.gov guidelines (HHS Web Communications Division, 2009) are often referenced relating to usability, they themselves being based on many usability works published over the last 50 years, but these are aimed at websites and online portals rather than specific mobile devices. However many of the ideas and guidelines provided can equally well be applied to situations beyond the original web based ones.

Gong & Tarasewich (2004) have produced a set of guidelines aimed specifically at creating usable interfaces on handheld devices. Many of the older guidelines are aimed at devices with large screen sizes (certainly larger than would be considered portable) and so the way information is presented is less reliant on having the space available to read on the screen. Large volumes of text for example are very difficult to display on mobile device screens without either having the text too small to be easily readable or having a significant amount of downwards scrolling to read all of it. The guidelines are broken down into two categories, firstly, those guidelines which are entirely new and appropriate to mobile devices only. Secondly, any guidelines which are modifications of already existing usability guidelines and have been adapted for handheld devices.

Some of the guidelines, although applicable to mobile devices, are still not applicable to authentication systems on mobile devices, particularly graphical ones. For example, providing word selection rather than requiring textual input from the user would not be secure in an authentication system. Also, the guidelines relating to how large volumes of information are presented by only showing the highest level of information and allowing the user to choose to see more detail would have little bearing for a login system.

Renaud (2009) presents a set of usability guidelines for graphical authentication systems explicitly, although for a generic graphical system rather than one tailored to mobile devices. The types of images which should be used with a graphical authentication method are specified, along with the feedback given to the user to ensure a usable system. Some security aspects are covered by these guidelines as well as the usability sides, such as requiring at least four pass images to make the system secure enough. Similarly for secure systems, or high risk systems, they suggest that users should not be allowed to choose their own images for the password since this makes them more guessable and there is a higher likelihood of the images choosing being distinctly different from the decoy sets of images. One other suggestion made is that on the web, pages should expire immediately, the picture names should be randomised and users should be able to easily change their passwords in case they become compromised. In terms of the usability of the system, the guidelines created are based on many of the same principals and concepts as the standard usability.gov web page guidelines (HHS Web Communications Division, 2009) such as providing helpful error messages in case of failures, progress indicators to allow people to see how far through the process of authentication they are and accessibility alternatives for deaf users. When choosing pictures, there ought to be no need for scrolling to see the entire set of picture. This is unlikely to be feasible on a mobile device due to the smaller

nature of the screen, making it impossible to show a large enough set of images at once without the images being far too small to see.

Miller (1956) looks at the ability of humans to remember more than one thing at once, concluding that the optimal number of bits of information that can be stored concurrently in the brain is seven. This has a large bearing on the usability of an authentication system, since if a system requires users to remember too much information simultaneously then they would not be able to cope and the system could not be considered usable.

4.3 Psychological Aspects

There has been a lot of work done into the psychology of human computer interactions. One significant area of research, Distributed Cognition, looks into the social or contextual aspects of HCI. This looks into how a user's position within their social or work group and their general mind-set can affect the way they use and perceive a system. This can be used in practical ways to present different but more intuitive ways of organising data, for example, by analysing the way users organise paper based information into the most accessible format.

With the advent of the internet in the early 1990s, work in HCI shifted from being predominantly based on office based software (word processors/spreadsheets) onto web based applications. Much of this work considers how users look for information both on a web page itself or how they sort through the mass of information generally available on the web. This analysis of behaviour can, in combination with design models intended to draw

on user experiences (for example User Centred Design), dramatically improve all aspects of the usability of a system.

This work is particularly relevant for mobile phones and devices. Due to the nature of the devices, the information displayed has to be short and to the point. However users will still expect basic functionality to conform to the standard ideas of UI design (e.g. Buttons are for clicking or, as in the case of a touch screen phone, pressing). This thinking is encapsulated in one of the criteria presented in Chapter Five stating that any design should use 'Standard Task Sequences' meaning that anything which the user sees ought to be intuitively obvious either about what it does or what it requires them to do. This should not even assume a basic familiarity with computer systems as a whole, just a general understanding of social concepts (red = bad, buttons are for pressing etc).

4.3.1 Recall vs. Recognition

When looking at authentication, the work load on the user is considered to be an aspect of usability, specifically the user satisfaction element and efficiency elements. As such, any authentication method that requires the user to remember too much information will be inherently unusable. Psychological studies have shown that people are better at remembering through recognition rather than through full recall (Hollingworth, 1913, Shepard, 1967, Tulving & Watkins, 1973). Recognition is the process whereby a person only has to choose an image or piece of information when shown it, whereas recall would be reproducing that information from scratch with no prompting. On a purely theoretical level, and discounting factors involved with the design of the system, it would make sense

that authentication methods relying on recognition ought to be more usable than those relying on recall alone.

There are however documented issues with graphical based passwords. Towhidi & Masrom (2009) highlight some of these through a survey of recognition based graphical authentication methods. They show that some of the main issues facing users are to do with choosing obvious or common images as the password, which makes them easy for an attacker to guess, as well as not being able to remember the sequence or order images were chosen over time (assuming the method in question requires the order to be preserved when logging in).

Authentication methods are based on one of three different types of memory: cued recall; pure recall or recognition. Lashkari et al. (2009) (extended in Masrom et al. (2009)) discuss a number of different methods from each category and well as the security features they provided inherently. They do not draw any conclusions about the differences between cued recall and pure recall methods, confining themselves to general observations about how people use the different methods. For instance, they find that users often choose weak passwords, and many users will choose the same or very similar passwords if a pattern is required to be drawn. For example, it might be expected that users who want a simple password would draw a straight line across the top of the drawing canvas which, although easy to remember, would be very insecure. They also conclude that the main issue in design relating to the usability of graphical authentication systems is that the researcher focuses too much on security forgetting the usability side, or vice versa. Thus a balance is needed between the two.

Weir et al. (2010) look into the usability of authentication methods for eBanking systems. They compare three different password systems across 141 users, a two layer password, and two onetime only pass codes (OTP) created either by a specific device or sent by text message on request. Whilst the OTP methods scored higher for security, people scored the 1 factor password higher probably because it is the method they were used to. This led to a slight preference for this type of password overall, although the results were otherwise evenly spread. They do demonstrate that the perception of authentication methods amongst the public is often wrong, particularly when related to methods they are already using. Many of the participants perceived the 1 factor authentication method already in use to be the most usable and secure, whereas the one time password method is in fact the more secure. Some inexperienced users even believed the OTP method to be less secure than the older password method. Overall it shows that people are unwilling to move to a new method of authentication with which they are unfamiliar, and although this paper deals mostly with eBanking issues, the reluctance to change is likely to be mirrored across most systems using an authentication method of some sort.

4.3.2 Dealing with Multiple Passwords

The issues with remembering secure passwords are compounded by the large number of passwords people are either suggested or required to remember. Many websites require a login in order for a person to use the features they offer, and if standard password practice is followed, no sites should share a password. With a large number of passwords this can become nearly impossible resulting in people often using the same, or very similar, passwords any time they are required to choose one. This has security implications in that if an attacker gains access to a password for one site, they would automatically be allowed

access into all of the other accounts owned by that person. As such, several methods have been proposed in order to allow users to create different passwords for each account they own in such a way as to allow them to easily remember them, or at least deal with having multiple passwords.

One of the main reasons that systems are insecure is through users finding the security too complex, and so finding ways around it. To combat this, users often need to be trained to use the system correctly, and so in the most secure manner. This also applies to the process of password creation, whereby a lot of users choose insecure passwords through lack of experience or knowledge of what a secure password is. Adams & Sasse (1999) look at several ways to create secure and usable text based passwords, as well as training users to create them whilst giving informative feedback about their choices. They also suggest that where possible, users should not be required to remember multiple passwords and at most four or five passwords if this is unavoidable. This particular recommendation is perhaps no longer relevant with the internet and the substantial number of sites requiring sign ups and logins all with different passwords. The security systems also need to be seen by the users as working, as well as describing why it is necessary. Plenty of feedback is again required to ensure users do not feel overwhelmed by complex systems which they would otherwise feel have no relevance to them. Similarly, only information relevant to them should be given to users, rather than presenting them with the security specifications of an entire system for which they only need to use one small part.

Mnemonics can also be used to help users to remember longer, and more secure passwords, by creating a story to go along with the password which is easier to retain. Users can be encouraged to create these on their own or they can be automatically

generated for random passwords (Jeyaraman & Topkara, 2005). Since the mnemonics are produced automatically, they could be sent to the user along with their new password. However, they do not test their system on real users, instead only showing whether their system can produce the mnemonics for random passwords.

Whether users can remember multiple graphical passwords will become more important over time as more authentication methods move towards graphical based approaches. Everitt et al. (2009) looks into users' abilities to remember and cope with multiple graphical passwords on the grounds that should they become more prevalent, users will start to experience similar problems with multiple graphical passwords they are currently facing alphanumeric ones. They find that even with as few as four graphical passwords to remember, users are more likely to fail authentication completely rather than eventually remember the password. This suggests that the increased usability of graphical passwords may be over estimated and that if they become more prevalent over time this effect would quickly be diminished. They suggest people have issues in this respect due to confusing the components of the graphical password in their heads, particularly if they are used infrequently. One suggestion they make to avoid this would be to make the user input their password often during an initial training period (maybe a week) and then eventually scale it back to a more appropriate level when they are more likely to have fixed the passwords in their memory.

However, work on graphical passwords at ATM machines has shown that people are able to remember up to five different graphical passwords relatively easily (Moncur & Leplâtre, 2007). They also tested whether mnemonics improved the performance of the graphical

password recall and found it did help. In either case recall performance was improved over that of a standard PIN.

4.3.3 Usable Passwords

As previously mentioned, the usability of passwords is a large issue when passwords are required to be too complex for users to remember. Again this can lead to security risks as people become tempted to write down their passwords allowing for the possibility that an attacker could steal them.

A lot of work has been put into this issue over several decades. For instance, Barton & Barton (1984) look into ways to generate passwords which are easier for users to remember, rather than the standard method of using a random string of characters. They suggest several ways of generating easy to remember passwords, such as using the first letters of each word of a line from a song or poem or some other similar phrase. These can produce complex passwords (although only alphanumeric ones) which are easy to remember but difficult for an attacker to guess. Similarly, using parts of words combined with other parts of other words can produce similarly complex non dictionary words to use as passwords. They also describe the use of methods of password creation using the environment surrounding the user's workstation, thus turning the recall of the password into a cued recall method relying partly on recognition (some of these methods are described in Section 3.1). As shown in Section 4.3.1, recognition has been shown to be easier for people to cope with than pure recall.

Eljetlawi & Ithnin (2008) look at the usability aspect of current graphical passwords and compare them to the currently accepted ISO usability standards (ISO/IEC, 1998, ISO/IEC, 1999) to see which if any of the methods is the most usable. They find none of the then current implementations fully fit into the ISO usability standard. This work was then updated in 2010 to include more recently released graphical authentication methods but with no difference in their final conclusions (Eljetlawi, 2010).

As previously mentioned, very secure passwords are often the hardest for people to remember, so training is often used to help people both create and remember complex secure passwords. Some websites require complex passwords and help to train users to create better passwords for use on their sites. Herzberg & Margulies (2012) present the results of a study which looked into the usability of sites which forced users to use secure passwords, and trained them to create better ones. They produce a variety of results on how users deal with issues such as spoofed login pages (specifically that users will be fooled a lot more easily if the page is linked off the homepage of a site they trust rather than from a random email). They follow this up by presenting their own authentication mechanism based on the results of their study. This is a graphical authentication mechanism which is meant to be learnable by users simply through use of the system, so needing no formal training. The method they developed is one of those discussed as a potential graphical authentication method in Section 3.1.3.7.

4.4 Mobile Usability

When applied to mobile devices, the usability of any system depends far more on the size of the information displayed than any other factor, due to the limited screen space available

with which to display information. Information therefore needs to be displayed in a concise format that still conveys the required meaning.

Many of the concepts required for mobile web browsing can be adapted or applied directly to mobile apps and these are at least partly taken into account with the design of the mobile security criteria in Chapter Five. For instance, Gong & Tarasewich (2004) present a series of guidelines to help in the creation of usable mobile interfaces basing their work on previous usability guidelines and updating them where necessary to be relevant in a mobile context.

It is now well established that people are better able to recognise images than they are able to remember strings of text (Shepard, 1967, Standing, 1973). Graphical authentication mechanisms have evolved since the first method proposed by Blonder, G.E. (1996) to take advantage of this fact and allow for theoretically more secure login systems which are easier for people to remember and use. An overview of some of the methods involved was provided in Section 3.1. However designing authentication systems for mobile devices produces a new set of challenges above those faced when designing authentication methods for desktop or other larger systems.

Mobile devices are innately smaller than their desktop counterparts both in terms of screen size, keyboard size (both virtual and physical) and for the moment, processing capability. The usability issues associated with the change in hardware need to be addressed before an authentication system can be successfully produced for a mobile device.

The small screen size will often mean that if a method requires choosing pictures from a larger set of images, then some form of scrolling will be required in order to view all of the images, since to make the images small enough to all fit onto the screen would mean they would be also too small to see. Whether or not users will actually make the effort to scroll through a large set of images would need to be investigated, as an overly long set of images would mean increased frustration for users who did scroll through them all. Those who didn't would miss a possibly important set of possibilities or options.

The keypad entry on a mobile device is often a virtual one as part of the touchscreen on modern smartphones. This further limits area of the screen available to the application since a portion of the screen will be taken up by the keypad if any text entry is required. Furthermore, when used on a touchscreen, if the QWERTY style layout is used, the letters used are small which can lead 'fat finger' problem where a user is not always able to see which part of the screen they are pressing with their finger. It has been shown there is no significant difference between younger and older people in terms of the ability to use buttons on a touchpad, there is however a large preference for larger buttons to be used by older people Siek et al. (2005).

Devices with a physical keyboard are often limited to a number pad with the digits 0-9 and characters as alternate functions of those keys. This leads to issues with inputting passwords where, without predictive text, a user would have to press significantly more keys than the length of their password in order to type it correctly. However, graphical passwords can be set up to allow keypads such as these to represent the images displayed as part of the password (assuming there are no more than 12 images displayed at any given

time). This not only gets around the issue of text based passwords but allows a graphical password to be used on a non touchscreen style phone.

How well users are able to locate functionality on their phones is another major issue with mobile devices. Without having shortcuts to everything visible on the screen at all times (impossible on a smart phone with limited screen size and a large number of apps), some form of menu structure is required, and these are often difficult to navigate and locate specific applications. Klockar et al. (2003) performed a statistically insignificant usability test on nine people with their phones to try to see if there was any indication of differences between different makes of phone and how easy it is generally to perform or find common functions. The main issues they found were users not being able to locate functions easily in the menus and the user's lack of knowledge of the difference between phone and SIM card memory, leading to filling up one and not understanding why nothing more can be added. The test itself was carried out by the user sitting using the phone and an observer watching their actions and counting the keystrokes required to perform said action. As such they propose a more simplified menu structure, based on a better understanding of a user's mental map of menus, or prioritised menus based on frequently used functions to overcome current difficulties.

Another way around the issues of menu structure is to use an intelligent application to learn which applications or options the user goes to most frequently. This can then be displayed in a more prominent position in the menu allowing quicker access to the most used functions. This would become personalised to each user over time rather than being those expected by the designer (Greene & Finnegan, 2003). A prototype version of this was produced, along with a similar system for sorting email messages based on subject

lines and user preferences as well as its past experiences of sorting mail from specific people or with specific subjects. This style of idea can be applied to many forms of applications so they are able to learn from the individual user and to a certain extent automatically tailor themselves to that user's needs. They plan to extend their system to include inputs form sensors on the phone or other devices which can be utilised to further improve the profile of the user. This would help the application more intelligently assess how to make itself more usable for that specific person.

4.5 Usability in Real World Environments

Several studies have been carried out into how security systems are actually implemented in real world situations. Often these show that despite the amount of literature and work into more usable security systems, current implementations often ignore the rules and as a consequence, produce often unusable or confusing applications.

For example, Smith (2003) looks at the human factors relating to security and HCI, and comments that even after a long period of having security systems in place, they are generally still not usable, citing issues with how cryptographic functions like PKI are not set up to match the mental model users actually have of securing emails. For those people who have grown up with these systems, the general idea is that security is an irritation to be avoided or sidestepped rather than embraced and used effectively, and for effective security to be implemented, this concept needs to be changed.

Another important paper in terms of usability history is related to specific issues with PGP 5.0, but applies to many security applications available at the time (Whitten & Tygar,

1999). It tries to explain why users are afraid or unable to use PGP security due to its overly complex nature, thus increasing the security risks to themselves by not encrypting important or sensitive information. They show that even though PGP 5 has a relatively good interface, it is not good enough for the average computer user to understand and implement. They use the method of cognitive walkthrough to evaluate the software whereby users approach the system whilst trying to act as if they were novice users. They also incorporate a heuristic set of targets for successful use of the system for a real measure rather than just finding possible areas of non usability. Only a third of the users taking part in the test, who were experienced email users, were able to properly send a secure email using the software. They cite issues such as the lack of obvious information about the importance of encryption as well as an overload of non standard functions confusing the user into which ones they need for sending a simple email. They suggest a simpler interface that would improve the usability of the product and a design based on the cognitive map of what users expect from the program.

It is expected that over time, security systems will become more usable as researchers and system designers create systems with a more user centered design philosophy. However research has shown that often, additional features or security updates can hinder the usability of a system by removing older features and adding new ones users cannot understand. Furnell (2007) looks at how security has evolved over time in two common desktop applications, MS Internet Explorer and MS Word. They find that although over time upgrades to the software have removed a lot of usability issues related to the security of the programs, the new features added seem to create new issues rather than having been designed from a usability standpoint, particularly in the case of Word. This is then

compared directly against Nielsen's Usability Heuristics (Nielsen, 2005) to show how many of the heuristics are not followed even by applications as common as Word.

Whether a system is accepted by a business is often dependent on how the users of the system see and use it over time, and many different factors can affect this (Gahtani & King, 1999). Although aimed at business environments, many of their hypotheses could be applied to a personal mobile device scenario. For instance, one of their hypotheses is that the enjoyment of using the system will impact on whether the system is accepted within the organisation, for which they find partial support. Another hypothesis, that the usability of the system would have an impact, was fully supported by their study.

This shows that if a system is particularly unusable, then people will tend not to use it to avoid the hassle and stress of not understanding what to do. In terms of a security system this can lead to users switching off security features, regardless of the increased risk this puts them under. This even applies to services where an a difficult system would often not be used by people at all, simply because it is so unfriendly, regardless of the features it may have (Markova & Aula, 2007).

This extends to purchases people make as well. Mack & Sharples (2009) look into how much usability affects users when choosing a product to buy. They find that although many users report usability as one of the main factors when deciding on a product, when it comes to actually buying, other factors such as the price and the aesthetics may be of more importance to them. The importance given to the usability factor will often change depending on the age, social standing, and gender of the particular user.

There is a significant difference in users' minds between what they perceive as a security risk and what actually is a security risk (Tari et al., 2006). Tari et al. (2006) look at the differences between the real and perceived risk of shoulder surfing style attacks when comparing alpha numeric and graphical passwords. The idea that an attacker could simply look over a user's shoulder and see the image or pattern being selected is a major hurdle in the introduction of graphical passwords and has in itself been the inspiration for several types of new graphical password methods to avoid this (see Chapter Two). They find that PassFaces is more vulnerable to shoulder surfing style attacks, partly due to the inherent nature that it is easier for the user to remember the password, but using the keyboard as the input mechanism rather than the mouse or number pad can remove this almost entirely. When looking at alphanumeric passwords, the authors found that non dictionary passwords were more at risk to shoulder surfing than dictionary ones. They suggest this may be due to the attacker having to concentrate on the characters of the password itself rather than being distracted by the meaning of a dictionary word if one is used. Here the perceived and actual risks differ significantly, with users expecting dictionary words to be more vulnerable to shoulder surfing attacks.

Overall however, it is shown that even alphanumeric passwords are more vulnerable to shoulder surfing than people expect them to be, particularly when compared with the perceived risk of graphical passwords. The mouse is the most vulnerable method tested for using PassFaces due to the nature of actively clicking onto the required image in a way which cannot be hidden. This has significant implications for touchscreen mobile devices where the user has to choose which images using a very similar method, rather than an abstracted keyboard press with no indication of which image was selected.

Payne & Edwards (2008) present a series of studies documenting previous work into password and other security usability issues. They begin with a survey of different types of passwords and their generally associated usability issues, many of which are to do with teaching users how to use them effectively. They also briefly look into the graphical password methods available at the time they wrote the paper. The studies are split into two sections, the first with two successful usability designs and the second with two unsuccessful ones. They use guidelines from Yee (2002) as a basis for whether or not the systems they review are secure. The first of the successful designs is called Salmon, and allows users to set file permissions easily for business networks. The second, Network in a Box, allows users to set up secure wireless networks by using gestures rather than having to set up wireless keys and security manually. Both of these are shown to pass all or nearly all of Yee's guidelines and are shown to score better on independent usability tests than the two worse designs.

They then go on to mention two studies representing bad usability designs. The first, Kazaa, does not give enough information to the user about how much of their computer is being shared with the rest of the Kazaa network. Specifically if they download to the C:\ drive then the entire C drive is shared, rather than just downloads, including emails, personal documents etc. The second, PGP, is shown to be too complicated for users to understand and the help given involves too complex terminology for average users to know what to do with it.

The capability to send secure and encrypted emails has been available in modern email clients for some time now. However, very few people actually use it, or even know it exists. A lot of this is due to the lack of usability of such features and a lack of

understanding by users of what the terminology used actually means. Most users who know of it will therefore avoid it, or switch the features off. Until these issues are fixed, and the information presented in an easier to understand format, only 'power users' will use the software to its full potential as they are less afraid of doing something wrong (Kapadia, 2007).

There are similar issues related to users creating their own ad hoc networks using a secure PKI infrastructure. Balfanz et al. (2004) look into how usable this is on a modern computer and show that most people would be unable to perform the task due to not understanding the meaning of information presented to them. From this, they develop five high level guidelines for developing usable security systems such as making sure usability is integrated from the beginning rather than retrofitting it after development has finished.

A similar method, Out of Bounds (OOB) communications, is also intended to enable secure communications between two mobile devices on an ad hoc basis. As with PKI networks, the usability of the setup procedure is the limiting factor in how practical these are (Kainda et al., 2009). It requires users to implement the channel and secure it by some form of preshared pass, which in physically close devices can simply be communicated from one user to the other. This type of communication can also be used to transmit RSA keys for later more secure communications. As OOB is public, it cannot be guaranteed that the information will not be 'overheard'. It does however add an extra level of complexity for the users themselves as they are required to manually check the integrity of whatever passphrase/key/image is being used for communication. They go onto compare the different methods for creating an OOB and show how certain methods are not secure enough. For example, compare and confirm, where the same string/word/thing is shown

on the screen of both devices and users click to confirm they are the same to authenticate each other. They also show that other methods, although secure enough, are unlikely to always concentrate fully on the process of authentication, which lessens the strength considerably. Other environmental issues will make OOB far less secure than it should be simply due to human error.

The MoFAX fax machine software for mobile devices allows mobiles to send documents stored directly to fax machines. Wright et al. (2005) look at this from a usability perspective to see whether it needed any work to make it more user friendly. They found many errors such as inconsistent menus, terminology and lack of confirmations all of which make the system unfriendly especially to novice or casual users. From this they then implemented their own version of MoFAX intended to solve some of the usability issues associated with the original.

4.5.1 Website Usability

Much of the work into usability over the past two decades has concentrated on the usability of web pages. Many of the principles are easily transferrable to desktop applications and mobile devices. Many of the criteria developed later in this thesis are based on the Usability.gov guidelines either in their original format or adapted specifically for use on mobile devices (HHS Web Communications Division, 2009).

One of the original methods for getting internet onto mobile devices was the Wireless Application Protocol (WAP). The usability of this was limited heavily by the small size of mobile screens at that time, and so the usability of presenting data was a vital part of how

pages for this were designed. However at the time, WAP web pages received a lot of negative reviews due to their complexity and how data was displayed (Buchanan et al., 2001). They suggest improvements to the usability that would make people feel more positive towards it through the use of a set of guidelines for WAP style data retrieval. This is mostly obsolete now particularly with regards to smartphones but some of the concepts explored are useful for displaying information briefly but without losing the required content.

On a similar note, Kaikkonen & Roto (2003) look into how users actually navigate pages on mobile devices over WAP where page content is reduced significantly to fit onto the device screen. They find that, unsurprisingly, some of the guidelines from standard usability texts do not apply to mobile XHTML pages, with the pages needing to be redesigned to suit their own usage. Again, with the advent of smartphones, this sort of research into WAP is partially obsolete but the guidelines they suggest are equally relevant to modern mobile web pages as they were to WAP ones. These include ensuring the user knows where they are at any given time by the page title, and by designing the page for the use of the back button rather than shortcuts back to the homepage.

Menu structures including a page title and how the information is delivered are another vital component of mobile internet connections, particularly given the small screen size (Kaikkonen & Roto, 2003). They discuss different methods of arranging information for access for mobile users concentrating mostly on web applications. They also present a set of guidelines detailing the sorts of things mobile users expect from their device which can be just as easily applied to mobile applications as they are to mobile websites.

Authentication on mobile devices is also different from a standard web browser due to the lack of a large keyboard for users to easily type in passwords, as well as the smaller screen. Halpert (2005) look specifically at the HCI issues facing authentication design on mobile devices. Their work is concentrated on the authentication process for websites on mobile devices, where the standard web usability rules have to be modified to be appropriate for a mobile sized screen. They also make a list of suggestions for the design of such pages, many of which can be easily applied to login screens on mobile devices, such as ensuring icons used to represent links or actions are easily distinguishable.

One of the main methods used on the web to confirm a request to a website is a human rather than automated script through the use of CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart). These are however primarily generated in English and can often be difficult for foreign language users to understand and use (Yan & El Ahmad, 2008). They also show that although many CAPTCHAs use colour in an attempt to increase the security of the image, this makes little or no difference other than to make the images more confusing for the user. Finally they propose their own framework to generate secure and usable CAPTCHAs avoiding the issues they found in their paper.

4.5.2 Usability of Mobile Devices in a Medical Environment

Within a medical environment, healthcare workers need access to the same set of patient records at a number of locations within the same hospital. The usability of such systems is paramount in these situations where the speed in locating the required information could have a significant impact on the health of the patient.

Since usability testing of devices cannot easily, or safely, be performed with critical equipment in a hospital environment, several studies have looked into how well these environments are replicated within the testing stages for new devices or equipment. Guidelines produced by Alsos & Dahl (2008) are intended to help researchers test whether their setups are realistic enough, along with showing how video and audio records of the test can enable a better analysis of any usability issues found. They later continue this by specifically looking into how mobile devices can be usability tested by creating mock ups of the environment in which the device will be used, and trying to accurately replicate real life scenarios (Dahl et al., 2010). Particularly in the case of hospitals, there are likely to be a significant number of distractions in the environment which could affect the usability of the system in ways that could not be modelled in a purely lab based usability evaluations.

4.6 Usability Testing

4.6.1 Generic Methods and Frameworks for Testing Usability

There are no set guidelines or standards for testing usability on mobile devices, with most developers choosing to test their applications on a subset of the ISO usability guidelines (ISO/IEC, 1998). Several suggestions for frameworks for testing whether or not applications or mobile devices are usable have been presented, and a selection of these is discussed below.

Several prototype frameworks already exist for testing the usability of mobile devices, and three of these are tested by Lim et al. (2006), and they show that none of the prototypes managed to find all of the usability issues on the device, and that real time methods were

required with users to discover them. This has significant implications for any usability study performed as it shows that certain types of usability testing will not find all of the related issues within a system, specifically what they call low fidelity prototyping techniques. However, even considering this, the authors note that these methods were helpful in uncovering usability issues.

In their paper, Fiotakis et al. (2009) come to a similar conclusion about the suitability of usability testing methods having looked at lab based, expert evaluations and field studies. Their results mirror most of the known work on previous usability studies in that lab experiments and expert evaluations are a low cost way of finding crucial usability flaws in a system. A field study, although time consuming, can often find issues the other methods may not discover and they suggest using a combination of all three approaches to ensure adequate usability testing.

This is mostly due to the fact that users in lab based conditions act differently to how they would in a real world scenario. As the mobility of a user increases, the distractions and competition for the user's attention will significantly rise, as will the number of positions and angles of use of the device (Gorlenko & Merrick, 2003). As such, any lab based test can never truly replicate the experience of performing the test in the real world, and many usability issues will be entirely missed if only lab based tests are relied upon (Duh et al., 2006). However, real world testing also has issues in terms of the number of distractions involved with carrying out the test (noise, lack of privacy, movement) which distracts from the issues of the device itself and places more emphasis on the effect of the user's environment at the time they were taking the test. Generally users were happier in a lab environment although as has been shown, not all issues can be detected under these

conditions. Several investigators have compared lab based experiments with real world or field testing (Kjeldskov & Stage, 2004, Jambon et al., 2007). Generally, they show that the lab based testing is not as thorough as real world testing and finds far fewer issues. Jambon et al. (2007) go on to suggest that videoing people in the lab based tests helps to find issues which would otherwise go unreported.

From the perspective of web development, the concept of a field test is less valid, although there are still many different methods to test the usability of a website. For example, heuristic usability testing (expert user testing), empirical testing and using eye tracking software to see which areas of a website a user is looking at. de Kock et al. (2009) compare heuristic testing and eye tracking to highlight the different types of information which can be gathered by each. Eye tracking testing will find issues to do with the effectiveness and efficiency of a site, and heuristic examination concentrates on specific usability errors. As such, both would be needed in order to fully evaluate a site's usability.

There is always a trade off in a system between its security and its usability. For the most part the more secure a system is then the less usable it is likely to be. Users however will only rarely be aware of this on a day to day basis, so Ben-Asher et al. (2009) created a system which allowed users to change this trade-off between usability and security and then simulated attacks on the system to see how users would react. In situations where an attack was more likely, users were found to change the security settings more often than in situations where fewer attacks occurred. Their experiment is intended to discover the way users behave in response to threats, rather than specifically design or model a new system.

4.6.2 Mobile Usability Testing

Movement tracking can be used on a mobile device to track the motions and options chosen by the user whilst they are using an application. A small camera is attached to the top of the device pointed at the screen, allowing for a complete and accurate record of exactly which options were chosen, along with the timings associated with those choices (Schusteritsch et al., 2007). This allows the researcher to have a complete record of all the actions taken by the user as well as all motions of the user's fingers over the interface whilst it is in use.

Another method which can obtain better usability data, rather than questionnaires filled in after the device has been used, is to keep a log of all of the actions performed by the user whilst they use the device to perform a task (Kim, 2010). This will keep a record of all of the choices made by the user, both successful and unsuccessful, which may not be recorded through the use of just a questionnaire or other data gathering tools afterwards. In Kim (2010)'s example, they implement a web lookup system for non English domain names, and used factors such as the time taken for each user to find the required information to rate the usability of the system.

This has led some researchers to try to find ways to improve lab based testing to try to make it more accurately reflect a real world scenario. Yuan Fu Qiu et al. (2006) look into creating a standard way for testing the usability of camera systems on mobile devices, which could in theory be extended to other functions on the device. They define four new basic aspects of usability based on a number of other sources where usability is tested, Perception, Control/Action, Learning/Memorization and Evaluative Feeling. They then

tested three different camera phones on the basis of their criteria, and were able to validate their criteria as an effective means of testing usability in the process. The solutions to the usability issues they present are for the most part specific to the particular mobile phone and camera software being used, but they recognise that this can be an indicator of the device as a whole, and a complete study would need to analyse the phone software itself in terms of their own criteria.

Several frameworks already exist for evaluating a mobile device in terms of its usability. The set of guidelines produced by Heo et al. (2009) is intended to avoid the issues of applying standard desktop based usability criteria to mobile devices where they may not always be relevant. The set of criteria can be individually tailored to a specific device from the larger set available. Another approach to this problem is to have a background process logging all of the actions by the user, including both correct and incorrect actions, to enable a researcher to later analyse where users make the most mistakes (Balagtas-Fernandez & Hussmann, 2009).

4.6.3 Design Issues

One of the issues with creating usable systems is that the systems are not always designed with usability in mind, preferring to introduce functionality into the system and worry about how people will, or can, use it later. This means the usability of the system is an afterthought and so any substantial changes cannot be made without redesigning the system as a whole. Many studies have shown that by including the end users during the design of the system, any usability problems can be found early in the design process and removed.

User centered design in some format is a required part of the design process if a final system or application is aiming to be usable. Smetters & Grinter (2002) discuss some of the difficulties in the design of usable security systems and propose their own methods for integrating usability in the design process. Rather than attempting to modify existing technology, they attempt to redesign the underlying security procedures, such as how authentication in ad hoc networks is managed, to allow for a more usable experience. In terms of the ad hoc networks they propose a new method whereby the initial connection and device details are sent over a separate wireless channel, which prepares the connection for the full key exchange process to take place on a different channel. This initial channel can be heavily restricted in order to only allow certain devices to access it and create the initial connection.

As already noted, the smaller screen size of a mobile device heavily impacts on the usability, particularly if there is a lot of information which needs to be conveyed to the user. There are several ways to get around this, such as using additional feedback to the user in the form of sound or vibration (Brewster, 2002). Brewster's work also confirmed other researchers' findings that testing within a lab environment for this type of usability study was inadequate since it was unable to reproduce the reduced usability and additional cognitive load involved in using the device accurately with sound outdoors or whilst walking. However the addition of sounds to the buttons on their devices allowed the buttons to be shrunk by half without a significant impact on usability, but without the sounds this decrease in button size has a significant negative impact.

4.7 Summary

This chapter has dealt with some of the literature relating to usability, both on and off mobile devices along with the standards set out by the ISO. Knowledge gained from this review, along with the literature on security from Chapters Two and Three, enabled creation of the set of usability and security criteria for mobile applications presented in the next chapter. Following the review of the graphical authentication methods currently available and application of the criteria, three mobile authentication methods were evaluated in terms of their usability and security. The design and format of this experiment, along with the criteria used are presented in Chapter Five, and the results in Chapter Six.

5 Chapter Five – Design of Experiment 1

This chapter describes an experiment (Experiment 1) to evaluate the usability of three graphical authentication methods. In the Systematic Mapping Study described in Chapter Two, there is shown to be a lack of research investigating the usability of security systems on mobile devices, including authentication methods. Some of these authentication methods are detailed in Chapter Three. This chapter describes the design of Experiment 1, which evaluates the usability of three of the authentication methods, Awase-E, DrawASecret and PassPoints. The rationale for choosing these three methods out of the available methods is discussed in Section 5.3.

Before presenting the design of Experiment 1, the formulation and use of the evaluation criteria are discussed, and each of the criteria described. The criteria were based on a combination of previous guidelines for both usability and security. The selected authentication methods (Awase-E, DrawASecret and PassPoints) can be evaluated against only some of the criteria in an objective manner, whilst against others the methods must be evaluated subjectively. Because of this, the experiment needed to obtain users' opinions on whether or not the criteria were satisfied. The design of Experiment 1 is detailed including the hypotheses for the experiment, and how data was collected in order to test these hypotheses. Finally, the details relating to how each of the chosen methods was implemented on an Android system is provided.

5.1 Criteria

A set of eleven criteria were formulated by combining the most appropriate elements of mobile usability, security and general usability criteria (Ahituv et al., 1987, Gong & Tarasewich, 2004, HHS Web Communications Division, 2009, Yee, 2002). Where guidelines from these sources were similar, these were combined and those not relevant were removed to produce the eleven criteria shown below. A more detailed explanation of which of the guidelines the criteria were constructed from follows the list of criteria.

C1. The most obvious route through the application should be the most secure.

C2. All objects available to the user should be clearly identified as to their purpose and outcomes, and 'Standard Task Sequences' should be used.

C3. Information displayed should be organised clearly with proper emphasis on the most important information.

C4. Colour should be used to enhance the application rather than as its sole method of displaying information.

C5. The application should 'feel' fast to the user. Security features that take too long to complete have been shown to be annoying for users who are then inclined to just turn it off.

C6. Care should be taken when designing the system to stop the user having to remember too many details at once since users only have a limited capacity for storing short term knowledge.

C7. Should a system error occur, it should not automatically allow access to the user, but should inform them of what has happened, and ideally if possible prevent the error in the first place.

C8. Where passwords or other authentication details are stored on the device (or server), the data should be encrypted in order to prevent an attacker who may gain access from simply reading off the passwords.

C9. In the event of multiple wrong passwords, the application should silently block further attempts at logging in, possibly for a set period of time. If appropriate or applicable, should there be multiple repeated failures, the device should be locked until the identity of the user can be confirmed by a service provider or other authority.

C10. The application should not indicate whether or not a login attempt has been successful until the user properly submits their details, no comments should be made on whether or not what has been input so far is valid.

C11. The application should allow for easy use on a mobile device, with the limitations (small screen size, touchscreen input etc) that this entails.

Using these criteria it should be possible to accurately evaluate an authentication method for a mobile device and show how usable it is. If a method can be shown to satisfy all eleven of these criteria then it may be considered to be usable and secure.

The methods can only be evaluated against some of the criteria (specifically C7, C8, C9, and C10) in a purely objective manner and so the methods can be designed with these criteria in mind to ensure they are satisfied Details of how the realisation of the methods chosen meet these objective criteria can be found in Section 5.3. For the other criteria, the methods need to be evaluated in a more subjective manner, and as such an experiment was required to evaluate whether participants feel the criteria have been satisfied for each of the methods.

5.1.1 C1 - The most obvious route through the application should be the most secure.

In the case where there are multiple methods of authentication (say for instance with and without SSL encryption) the default choice should be the most secure. This could partly be said to be an objective criterion based on what the default choices on the system are, but it also applies to the design of the application and how it is used, since what is obvious to one person as the first choice might not be to others. This was adapted from (Yee, 2002) design principals, specifically the 'Path of Least Resistance' principle. This states that the most comfortable way to perform a task ought to allow the least authority on the system, ensuring users are required to do something to be authenticated.

The methods can be evaluated against C1 both objectively or subjectively, depending on the application in question. However for each of the methods built, there is only one route through the applications so by default it is the most secure, and so this criterion was not included as one of the criteria used in this usability study.

5.1.2 C2 - All objects available to the user should be clearly identified as to their purpose and outcomes, and 'Standard Task Sequences' should be used.

This criterion is built on the principal that all actions or tasks available to the user do what the user would expect them to, as well as looking for information in the places they expect it to be. For instance, menus and other options are expected to be either on the left or at the top of the screen and placing them in another position would confuse users, 2.3 (HHS Web Communications Division, 2009). Guideline 13.5 (data entry fields must be labelled) and the 'Identifiability' principle from Yee (2002), (all objects displayed must be easily identifiable relating to what they are or refer) combine to make this criterion as a whole.

This criterion is subjective since, similar to C1, what is obvious to the researcher may not be as obvious to a casual computer user. As such, the methods can only be evaluated against it through the use of a questionnaire after the system or systems have been used.

5.1.3 C3 - Information displayed should be organised clearly with proper emphasis on the most important information.

This criterion is taken from a combination of guidelines 6.1, 6.6, 11.10 and 16.1 from the Design and Web Usability guidelines (HHS Web Communications Division, 2009). 6.1, Avoid Cluttered Displays', and 6.6, 'Optimise Display Density' relate to not overloading the user with information on a screen since they would not be able to take it all in. 11.10, 'Emphasize importance' and 16.1, 'Organise information clearly' relate to the information that is still displayed on the screen and making it obvious which bits are the most important or relevant. This is more of an issue on mobile devices over standard websites due to the substantially smaller area on a mobile screen which can be used to display information.

This criterion is entirely subjective though in terms of how much each user feels would count as information overload. Although care can be taken in the design to minimise this issue, it can only be shown to have been effective through discovering the user's opinions of the system via the questionnaire.

5.1.4 C4 - Colour should be used to enhance the application rather than as its sole method of displaying information.

This is a combination of two criteria from the web and usability guidelines (HHS Web Communications Division, 2009). Guideline 3.3 states that colour alone cannot be used to convey information on a page, due to the relatively higher percentage of colour blind people in the world (on average ~10% of any given population). However, guideline 11.9

states that if information can be colour coded then it is more understandable. Combining these two leads would result in colour being used to present information more effectively, but only when there is a text based alternative. To some extent, different shading could be used to show differences in information since, although colour blind people cannot tell certain colours apart they can tell the difference in the lightness/darkness of a particular colour. In cases where the user is not colour blind, colour coding has been shown to significantly increase the speed at which users can absorb information on a page, when compared with the same data in text format. This is partially subjective as it can be designed into the system, but whether it is effective for colour blind people would depend on the individual person.

5.1.5 C5 - The application should 'feel' fast to the user. Security features that take too long to complete have been shown to be annoying for users who are then inclined to just turn it off.

Research has shown that when a security system is slow or inconvenient, users are inclined to turn it off altogether rather than wait (Balfanz et al., 2004). As such a system is better designed for performance rather than to specific user's preferences if a choice must be made between the two, which is the basis of guideline 1.6 from (HHS Web Communications Division, 2009). If a computationally intensive process is being carried out then the user needs to be informed that they have to wait before moving on with the task. Doing so should reduce frustration at having a system that randomly 'hangs' without giving any feedback as to whether it may have crashed entirely or is simply doing

something that takes a while. Gong & Tarasewich (2004) also present this as one of their rules relating to mobile interface design 'Application should be up and running quickly'.

This is more of an issue on mobile devices since the processing power available can be relatively low compared to desktop machines which users might be more accustomed to. However the advent of affordable smartphones this is becoming less of an issue. However, where possible computationally intensive processes should not be run on the mobile device itself due to the decrease in battery life this would cause. The speed of the system can be looked at both in terms of the amount of time taken to perform set tasks (for instance the authentication process itself after the user has entered their details) or the general responsiveness of the application (for instance the time taken for a button press to be registered and the response displayed to the user).

This criterion is entirely subjective, and each of the methods can only be evaluated against it through the use of a questionnaire to find out whether users felt the system or application was fast or responsive enough for them.

5.1.6 C6 - Care should be taken when designing the system to stop the user having to remember too many details at once since users only have a limited capacity for storing short term knowledge.

This criterion is aimed at reducing the memory load on the user whilst they authenticate themselves and would apply equally to all authentication methods on mobile devices and desktops, other computers etc. (Gong & Tarasewich, 2004). They go as far as stating that recognition should be used rather than memorisation for a generic mobile application or

website. Section 2.5 of the web usability guidelines states that users should not be required to remember information from one part of a website to another.

In terms of authentication systems this means that if there are multiple stages to the process, later stages should not rely on the information gathered during the initial stages (so all the information required in order for the system to authenticate the user should be given only once). Also the application should not require the user to remember too much information in order to log in, or they would not be able to remember it all. This would be partly removed by more constant use of the password in question, regardless of its length. However initially users might be inclined to physically write down passwords, or record them in another insecure way, in order to remember them better.

This criterion is mostly subjective in terms of whether users feel there was too much information to remember. However, there is an objective portion of the design to ensure that information is not required to be carried over between different sections of the application.

5.1.7 C7 - Should a system error occur, it should not automatically allow access to the user, but should inform them of what has happened, and ideally if possible prevent the error in the first place.

This criterion is based on Yee's principle of Explicit Authorization (Yee, 2002) which states that any authorisation gained in the system must have been granted specifically rather than by default. This is an objective criterion so can be built into the applications themselves. Specifically they applications should be designed so that the only way a user is able to be logged into the system is if the user has successfully entered their password.

Similarly if the application crashes, this should not allow the user into the system at all if possible.

5.1.8 C8 - Where passwords or other authentication details are stored on the device (or server), the data should be encrypted in order to prevent an attacker who may gain access from simply reading off the passwords.

All data stored on the device relating to the user's password has to be encrypted. This stops an attacker who has gained direct access to the database, for example through theft, from simply reading off a plain text password and so gaining access to the system. This is taken from the security guidelines for authentication systems by Ahituv et al. (1987). As with C7, this is a purely objective criterion and so can be built into the application. Hash functions serve this purpose well since they offer no inbuilt decryption method so a brute force approach would be required to convert the stored string back into the original text.

5.1.9 C9 - In the event of multiple wrong passwords, the application should silently block further attempts at logging in, possibly for a set period of time. If appropriate or applicable, should there be multiple repeated failures, the device should be locked until the identity of the user can be confirmed by a service provider or other authority.

This criterion is harder to implement on a testing application or system since it would be best implemented by the provider of a mobile server (or in the case of more stationary machines, the network administrator). It is taken from Ahituv et al. (1987) and intended to stop brute force attacks on the authentication system from succeeding. Particularly in the case with current PIN unlock systems on phones, this is already implemented on the majority of networks if the user enters an incorrect PIN too many times.

The current solution implemented by android phones with either a graphical or numeric/PIN unlock method is to prevent the user from entering a new attempt after 5 unsuccessful ones. If the user then continually tries and fails to unlock the screen, the device is locked entirely, and the user is required to enter the Google account username and password associated with the device in order to get access to change the passcode. The initial stages of this could be implemented on a device relatively easily by disallowing repeated fast attempts at logging into the system.

This is again a purely objective criterion so would be designed into the system. With the methods being evaluated here it has not been implemented as there is no way of sending the data to any form of administrator due to the absence of a working SIM card. This may not be implementable at all in some situations, for instance if the device does not have a mobile internet connection or subscription there may not be a definitive authority to unlock the device. One possibility might be to send an email to a predetermined address with an unlock code. However again this would require an internet connection on the phone to verify the entered code.

5.1.10 C10 - The application should not indicate whether or not a login attempt has been successful until the user properly submits their details, no comments should be made on whether or not what has been input so far is valid.

This criterion is also taken from Ahituv et al. (1987) and relates to failed attempts to attack the system. If the login process is a multi staged one then the user should not be given any indication as to the success of the intermediate steps until they finally log in. This is most applicable to the Awase-E method where a user is presented with a series of screens and is required to choose images on each of the screens. This stops an attacker from being able to

guess which parts are correct and false by giving them no feedback on the individual sections beyond whether they combined to make a successful or unsuccessful login.

This is another entirely objective criterion and so can be built into the applications. However, for the purposes of evaluating the methods, this was disabled to allow participants as many attempts as they like at logging in. This allowed analysis of the results to show how many attempts on average it took participants and so allow a reasonable maximum to be set in any final versions of the application.

5.1.11 C11 - The application should allow for easy use on a mobile device, with the limitations (small screen size, touchscreen input etc) that this entails.

The final criterion is taken from Gong & Tarasewich (2004) and relates to the use of applications in general on mobile devices and the associated hardware limitations involved. Particularly this relates to the small screen on a mobile device and the reduced amount of information that can be conveyed through it at any given time. The user should not have to scroll at all if possible, although in the case of lists of data this may be unavoidable, since as the inclination would be to only use the data in the first part of the screen, they would never see the latter part. During the experiment this was actually shown to have happened when looking at the images used for the Awase-E method (see Section 6.3.4.1). In addition to this, the usability of the application would decrease due to the extra time required in order to scroll or read all of the information.

This is partly subjective in that care can be taken when designing the systems to ensure that only one screen of information is displayed, but whether the application is considered usable on the smaller screen still depends on the opinion of the individual user.

5.2 Experiment Aims and Process

Three different graphical authentication methods were chosen for usability evaluation: Awase-E, DrawASecret and PassPoints. These three were chosen since they represent the three main types of graphical authentication methods available based on the type of memory required to use them: recall (DrawASecret), cued recall (PassPoints) and recognition (Awase-E) (Biddle et al., 2011), and are discussed in more detail in Section 5.3. As discussed in Chapter Four, recognition based methods ought to be easier for people to use because they rely on recognition memory rather than recall (Tulving & Watkins, 1973).

Awase-E uses a strategy based on recognising images previously chosen as the password. DrawASecret uses a strategy based on memorising a pattern and so can be categorised as a recall based method. PassPoints uses a combination of between the recognition (of areas on and image) and recall (f which areas were chosen as the password).

Experiment 1 was carried out to evaluate these methods against subjective criteria (C2-C6 & C11) from those described in Section 5.1. Participants were required to create an account and login using each of the methods on a mobile device, and a questionnaire was subsequently used to collect their opinions of the methods. To avoid bias, the same questionnaire was used for each of methods.

Although the C1 criterion is partially subjective when applied generally, it was not considered subjective for this experiment since for all three of the methods there was only

one route through the login process. As such it is purely objective and so does not need to be evaluated in the experiment.

Since all of the criteria which deal with the security side of the authentication methods were deemed to be objective, this experiment only considers the usability aspect of the methods. The security aspects are dealt with in the implementation stage by ensuring that as far as possible the methods meet the security criteria defined.

5.2.1 Hypotheses

The main hypothesis tested in Experiment 1 is as follows.

> *H01) The null hypothesis is that there is no difference in usability between the three methods being evaluated. The alternative hypothesis (HA1) is that there are significant differences in the usability among the three methods. Theoretically we would expect the recognition based method to perform better than the recall based methods. We would also expect cued recall to perform better than unassisted recall.*

There are also two ancillary hypotheses that have influenced the choice of experimental design.

> *H02) The null hypothesis is that the technical ability of the participants does not affect task performance using the three authentication methods. The alternative hypothesis (HA2) is that participants with a better understanding of current mobile*

devices background will perform better at the experimental tasks than those with little or no knowledge and will therefore score the usability of the methods higher. Theoretically we would expect computer and technologically literate participants to find the authentication methods easier than more novice users. To avoid bias due to the technology expertise of participants and to assess the impact of technology knowledge, we selected participants from two separate groups, one representing computer literate participants and the other representing novice users.

H03) The null hypothesis is that the order in which methods are used will not affect the usability scores. The alternative hypothesis (HA3) is that the order in which the methods are evaluated will affect the usability scores given by the participants where a method evaluated first or last would score better than a method evaluated second. Theoretically we would expect the alternative hypothesis to be supported, since many studies have confirmed the existence of such 'recency' effects in memory based tasks (Deese & Kaufman, 1957). The cross over design has been used to ensure that if such effects occur they will not bias any of the results.

5.2.2 Population Selection

Overall, 36 participants took part in the experiment. Half of these were University students aged between 18 and 21, referred to as Group A, and the other half were adults who were not students aged over 21, referred to as Group B.

The sample of participants in Group A was taken from students studying computer science at Keele University who volunteered to take part in the research. No incentive was offered to although they were encouraged to take part to gain experience of participating in such a study. The technical competence of this group was expected to be higher than average for their age group due to their area of study.

A convenience sampling method was used to select participants for Group B, by using members of the public who were receiving computing support from the researcher. This does introduce a selection bias from Group B, however, it also means that those selected are likely to be less computer literate than a random sample of the population (since they need help to solve their computing problems). This enabled the experiment to show how usable the methods are for participants with a limited level of computing literacy.

5.2.3 Dependent and Independent Variables

In this experiment, in order to test the hypotheses, the main dependent variable measured was the usability of the authentication method on a mobile device. The independent variable being used to test this is the specific authentication method (i.e. Awase-E, DrawASecret or PassPoints). Normally the device being used would be another independent variable since factors such as the screen size, or type of touch screen (whether it has a stylus for example) would have an effect on the usability. For this experiment however, the same device is used throughout as a control to ensure that the device in use does not affect the usability of the different methods.

In terms of the other observed variables in the experiment, the most obvious one is the technological capability of the participants. The participants were split into two groups to

determine whether the level of computing literacy affected the usability of the methods. Another of the observed variables is the order in which the participants used each of the three methods. This order was rotated for each of the participants so as to balance out the effect of learning how to use the device overall, however, the effect of this learning can still be calculated using the statistical method from Senn (2002) and is discussed in Section 6.2.

5.2.4 Experimental Process

Each of the participants was required to use each of the three methods making this a *within-subject* study design. Participants from each group were split evenly into one of six sets, and the order they use the methods was rotated depending on which set they were assigned to, as shown in Table 6. This was done to ensure an equal number of participants used the methods in each of the different possible orders. So, in each group there were 6 sets (a to f) of three participants each, who used the methods in the same order.

Set	Method Order		
	Awase-E	DrawASecret	PassPoints
a	1	2	3
b	1	3	2
c	2	1	3
d	2	3	1
e	3	1	2
f	3	2	1

Table 6: Order assignment of sets of participants to authentication methods

Participants were first presented with a set of brief instructions on how to use the different methods on the device itself before the treatments began. This was required since not only are the participants likely to be unfamiliar with graphical authentication methods, but they may also have little or no experience of using smart phones, or in some cases of using mobile devices at all. This was particularly the case for participants in the Group B, who were older than those in Group A and not as familiar with the technology.

In this situation, a treatment is defined as the participant using one of the authentication methods. As part of this, participants were required to create an account password using each of the methods in the order assigned to them. Following this there was a short break during which they were asked to perform some other activity unrelated to the experiment, before returning and attempting to remember the passwords they created in order to log on. After attempting to login using one of the methods, the participants were asked to complete the questionnaire to record their views on the method used. The login step was repeated for each of the methods in the assigned order. A diagram of this process is shown in Figure 20.

In both the account creation stage and the login stage, the participant was presented with the device in the correct state for them to perform the action required. During the entire process of using the device, the researcher was available to help, should the participant get too confused or stuck as to what to do next, but otherwise no assistance was offered. In cases where help was required, this was recorded to aid the later analysis of the data.

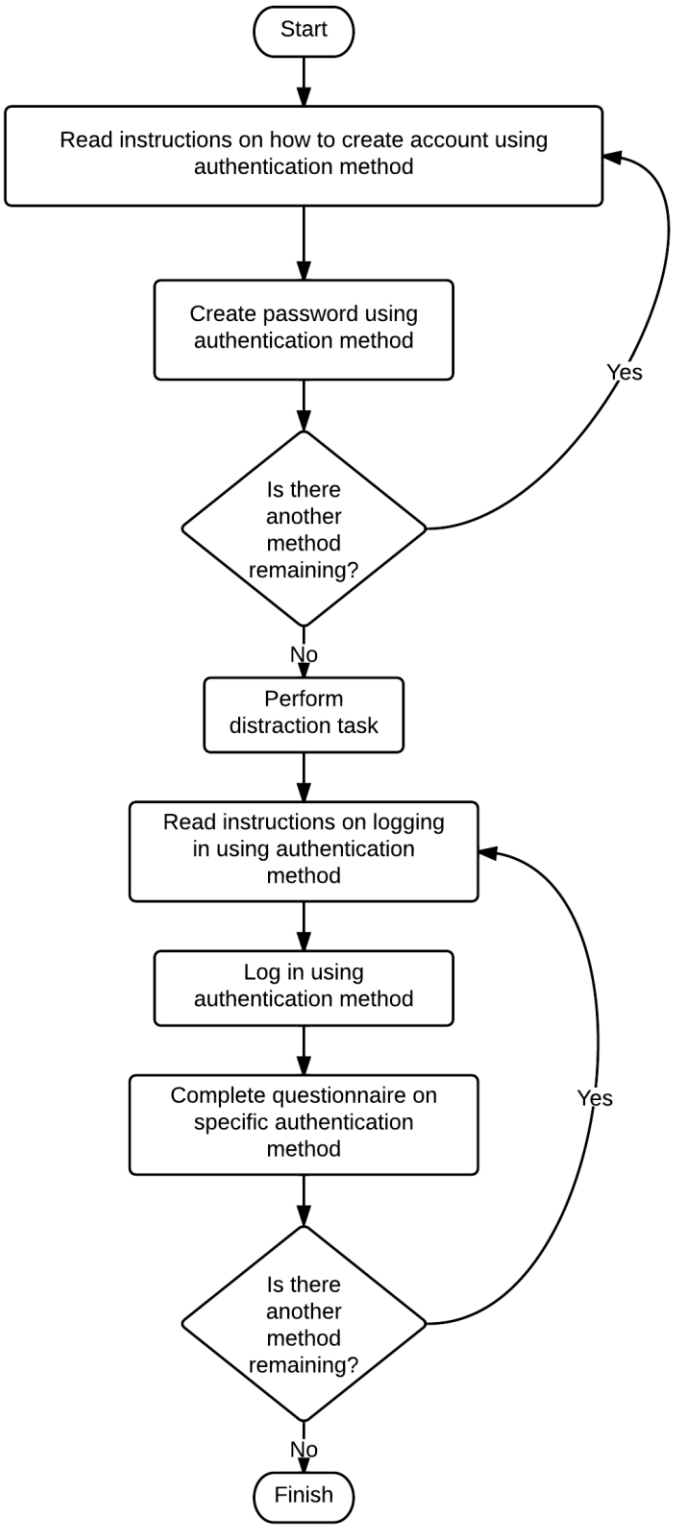


Figure 20: Process diagram of the experimental process as seen by each participant

To help minimise the effect of learning, the participants used the three methods in a rotating order to prevent the same method from always being the last to be used. This prevents the final method from consistently performing better due to the participants becoming more familiar with a touchscreen device, particularly in the cases where the participants have not used one previously.

With any experiment based on participant's memory, if the recollection part of the treatment occurs too soon after the creation process, then the participants would be easily able to remember their chosen password. This could lead to the participants giving better scores than they would otherwise, as the ease of remembering the password might lead them to forget any usability issues they experienced whilst using the methods. This was the reason for the short break in between the creation and login stages. In the case of the

students in Group A, an artificial time delay was introduced, often whilst another user was participating, so the time taken whilst other students evaluated the system helped to reduce any bias due to memory. For those in Group B, another unrelated activity was performed depending on the individual participant. During the study, several types of data were collected. For the purposes of collecting data, but retaining anonymity, each user was assigned a unique identifier.

Dataset 1 – Questionnaire

Data was collected using a questionnaire (see Appendix A) to obtain details of the participant's opinion of each of the methods, as well as any preference they might have about using each of the methods on their own phone. The questionnaire was originally based on the IBM computer usability satisfaction questionnaire (Lewis, 1995). Each of the questions asked the participant to give a scored based on a Likert scale from 1 to 7, as well as allowing the participant to add any additional comments should they wish to. Such additional comments were not however required.

Questions that were not relevant to either the methods being evaluated, or the use of a mobile device were removed. For the purposes of the questionnaire, the term 'application' was used rather than method to describe the methods. It was felt that this would make more sense to the participants in the context of a mobile device where most would at least be familiar with the concept of an app or application even if they had never used one before.

The final mapping of the questions on the questionnaire to the criteria is shown in Table 7. Additional questions relating to specific criteria were added, and some of the questions were reversed (so that a score of 7 on one was a positive score, but on the next this meant a bad score) to ensure participants were not just choosing the box on the left hand side all the way through. The questionnaire was filled out three times by each participant, once for each of the methods attempted. Only the subjectively testable criteria were evaluated over the course of the experiment, and so criteria C1 and C7-C10 are not included in this mapping.

Criteria	Questions
C2	1
C3	2, 3
C4	4
C5	5, 6
C6	7
C11	8
Gen1	9-11
Gen2	12-13

Table 7: Each of the subjective criteria matched to questions from questionnaire for Dataset 1

Some of the questions on the questionnaire did not map well onto any of the usability criteria formulated. These were the questions which related more to the user satisfaction element of usability, whereas the criteria focused more on aspects of efficiency or effectiveness. Consequently, two additional criteria were added to the original six being used in the experiment, named Gen1 and Gen2. These related to the final five questions on

the questionnaire. Questions 9-11 were combined into Gen1, and were intended to discover whether the participant enjoyed using the method. Questions 12 and 13 were combined into Gen2 and were aimed at discovering whether the participant felt the method was suitable for use as an authentication method on a mobile device.

Dataset 2 – Device Logging

Whilst using the device, each of the actions performed by the user was logged. This included all key presses, along with gestures and movements made on the screen, the intended action and the timestamp. An exact record of the chosen password was also recorded, along with any failed attempts at logging in as a whole.

In addition to the basic button presses, the type of data gathered for each of the methods varied. In terms of the Awase-E method, this involved storing a record of all the images actually selected on the 'Choose Images' screens including any that were deselected. All failed attempts at logging in were recorded along with the choice of images used in the attempt.

For the DrawASecret method, all movements made by the participant on the screen were recorded, enabling a complete reconstruction of the actual drawing made as the password, rather than just the squares the user passed over.

Finally, for the PassPoints method, all the points chosen were recorded, along with the squares to which those points equated, as well as any failed attempts at confirming the

PassPoints when creating the account. Similarly all the points chosen when logging in were recorded along with whether or not they matched with one of the PassPoints stored.

The logging of this data allowed for the identification of three distinct variables:

- The time it took participants to set up a password
- The time it took participants log in using their own password
- The number of times the participants input the wrong password

Dataset 3 – Qualitative Data & Observations

In addition to the data being logged by the device, the actions, and any verbal communication directly related to the creating account or login process were all recorded manually by the observer. This involved any requests for assistance in using the device, or application, as well as any actions performed that would not otherwise specifically be logged, such as the reasons why they were choosing the patterns or images they did.

5.2.5 Experimental Limitations

No laboratory experiment can ever completely accurately reproduce the conditions a normal user would experience when trying to login on a phone in real world situations, so this is an obvious limitation of the study. Furthermore, participants may feel pressured into answering in the way they feel they ought to answer, or how they think would most help the experiment, rather than with their own opinions. Experimenter bias is less of an issue with this experiment than others since no direct questioning of the participants took place. Development of questions for the questionnaire which do not lead the participants to

answer in a specific way minimised any effect this could have had. A more comprehensive look at the threats to the validity of the experiment can be found in Section 6.6. This section details threats to validity both from issues known before the experiment was undertaken, as well as threats discovered during the conduct.

5.2.6 Expectations

It was expected that the Awase-E method, which is based on recognition rather than full recall, would be the most usable. Even in terms of pure recall, Awase-E also is fairly low in terms of bits of information for each participant to recall (just four images rather than the almost infinite possibilities of DrawASecret or at least five locations and a picture for PassPoints).

For DrawASecret, it was expected that participants would have some difficulty when attempting to draw curved lines, particularly when they were over gridlines. Due to the nature of the method, curved lines will often pass close to the grid lines and could make the exact reproduction difficult if the drawing touched into another square by accident.

It was also expected that people who enjoy or favour the PassPoints method would be those who are good at memory associations. The concept behind PassPoints is remembering areas on a picture, rather than particular images or patterns. As such, people who are good at associating things in order to better remember them should be able to use this approach more effectively (for instance, remembering to choose the wheels of a car as one or more of the points). However this is also expected to lead to a security issue in that most people are likely to pick similar points if they choose the same pictures. It was

expected that an analysis of the points chosen by participants would confirm this (see Section 6.3.4.3).

The experience a user has already with touchscreen phones was expected to influence their ability to use the methods effectively. This was already partly accounted for in the separation of the two groups of users, where participants from Group A were expected to perform better than those from Group B due to their increased exposure and familiarity with technology in general. However within this, different levels of experience (particularly owning a touchscreen phone) will affect the participant's learning curve. As such the data on whether each participant had used or had experience of a touchscreen/smartphone was collected, along with whether or not they used any form of PIN, be it graphical or text based, on it. Any participants with a Google phone and making use of the graphical unlock feature ought to perform significantly better on the DrawASecret method since the two are similar in concept.

5.2.7 Limitations

The Android operating system has a feature in that from any given application it is always possible to press the 'Home' button to return to the home screen. In terms of an authentication application, this would obviously cause the application to be useless if it were implemented as a lock screen for the entire phone without it being built into Android itself. However, the application could still be used as an authentication method to protect other applications, such as payment or banking related ones.

There are already several examples of additional security or password applications which can be set to be required to be run before the target application can be opened (CarrotApp,

, Android Protector, SebastianApps). These come in the form of an extra level of security which can be attached to any app of the user's choice to prevent access to that app without the additional password. In this scenario, if the user, or an attacker, pressed the home button, they would merely be returned to the home screen and the target application would remain unopened rather than simply closing the extra authentication application.

The other buttons on the Android screen can be set to give no response, or return the user to the first page of the application. This prevents the application being closed entirely and possibly allowing access into the required security application on the device. For instance the 'Back' button would normally also return the user to the home screen if they are on the main or first screen of an application, however, this has been disabled. This ensures that in these eventualities, the user is not accidently permitted access through the authentication procedure, as is a requirement of criterion C7.

5.3 Implementation details relating to the three chosen methods

The three methods chosen were selected from the methods described in Section 3.1 on graphical authentication methods. Those could better be defined as strategies for completing the authentication, but for clarity they are referred to as methods throughout. Similarly, the actual 'methods' developed for Experiment 1 are the realisation and implementation of the original strategy, but again for simplicity, these are referred to as methods for the purposes of the experiment. A chart showing this is shown as Figure 21.

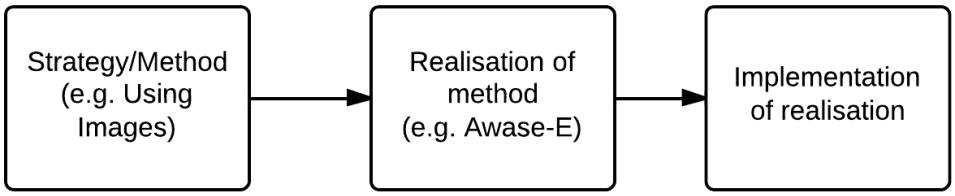


Figure 21: Breakdown of terms used to describe implementation of the original methods/strategies

All of the methods were developed for a mobile phone running using the Android operating system. Their design and implementation were based on the details provided in the original publications which proposed the methods. In creating these specific implementations, some small details of the methods had to be altered either to fit in with the constraints of the environment, or of the smaller screen of a mobile device. These details are provided in the following sections along with details about how the different implementations meet the objective criteria which are not being assessed as part of the usability study.

5.3.1 Awase-E

In this realisation of Awase-E, it was decided to use four images as the required number of images each user had to remember since this was the number in use in the demonstration version of the method created by the developers and created using PHP.

When creating the account password, it was decided that the Android gallery widget would be the most effective way of showing the large selection of images available to the user. There would be no way of properly showing all of the 100+ images available on the same screen so there had to be some form of scrolling involved, and the gallery method made it clear that it was a different step to the logging in procedure where the images were tiled.

An example of the gallery widget used in Awase-E is shown in Figure 22 and the UML activity diagram for the create account stage is shown in Figure 23.

Figure 22: Create account stage of the Awase-E method

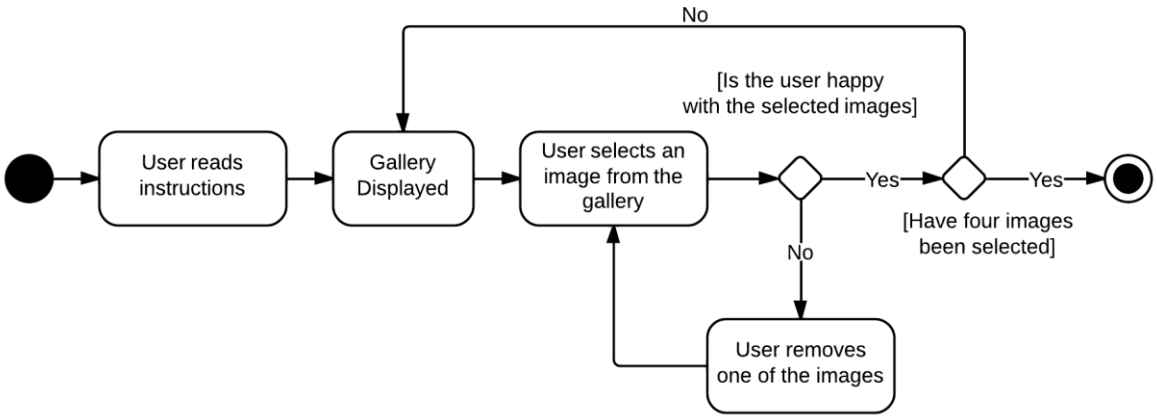


Figure 23: UML Activity Diagram for the Create Account phase of the Awase-E method

When logging in, the participants were presented with 12 tiled images (in a 4x3 pattern), along with the option for 'No Images' beneath. This is an increase over the number of images used in the original Awase-E specification since it was felt that this improved the security of the method (specifically to above that of a standard PIN code) without making the images too small to be recognisable. An example of how the 12 images are displayed to the user is shown in Figure 24 and the UML activity diagram in Figure 25.

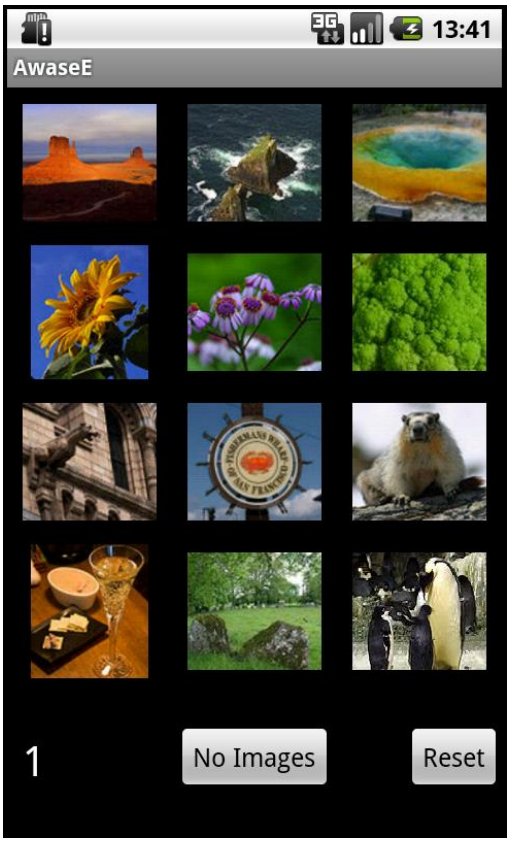


Figure 24: Login stage of the Awase-E method

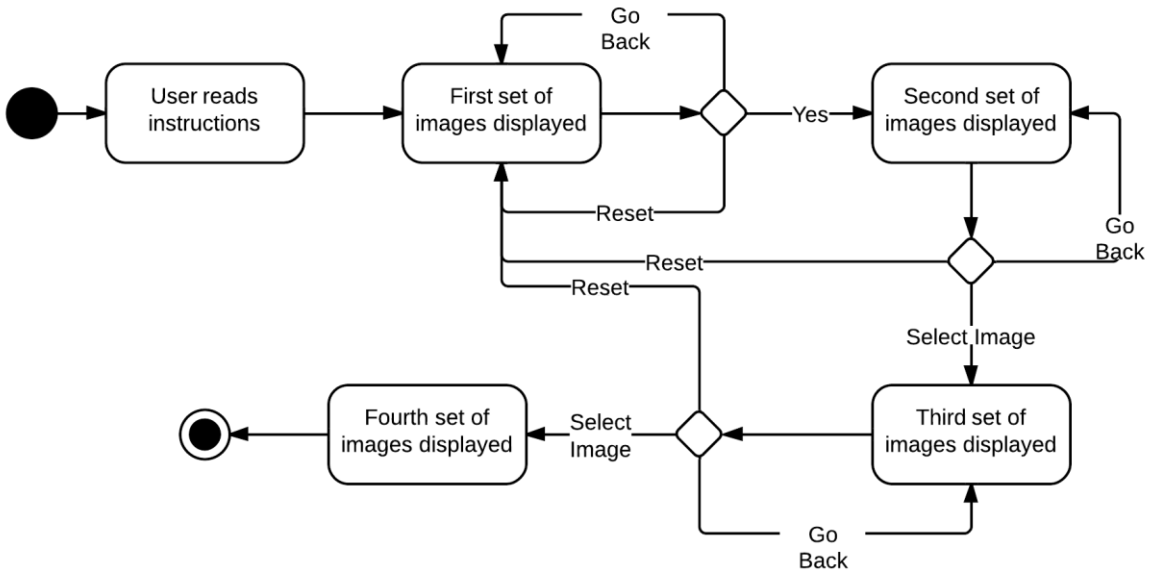


Figure 25: UML Activity Diagram for the Login phase of the Awase-E method

It is only possibly for users to choose from a set of images already stored on the phone, and they are unable to upload either their own images or images from the internet. In a full scale implementation this would be a distinct possibility although it produces a number of its own security risks, particularly in the way that any user uploaded images as these would be likely to be substantially different to the decoy set and so stand out more.

The probability of the No Images button being the correct choice on each screen is 1/10. This decreases the chance of an attacker guessing the pass images from repeated attempts at logging in since the images are slightly less likely to appear on each screen (1/10 vs. 1/13). However from a simple random guessing attack, an attacker has a 0.0001 probability of logging in by continually hitting the 'No Images' button compared with 0.000032 when picking random images. This means that if an attacker only used the no images option, they are three times more likely to get a false positive login than if they continually chose random images.

Objective Criteria

C1 is met in the implementation in this case by only having one login method available to the user which will automatically be the most secure option.

C7 is satisfied through code which catches any system error and in such an event returns the user to the main login/create account page. The error that occurred is then logged as far as is possible (for instance SQL errors could not be logged to the SQL file for obvious reasons) and displayed to the user, mostly for debugging purposes at this stage. In any full implementation of the application, the user would be returned to the login screen only, whereas for the purposes of the experiment, they were returned to the page allowing them to either create a new account or log in.

For C8, the data stored on the device in terms of each user's password are the names of the image files used. Each of the filenames stored within the SQLite database is hashed using an SHA-512 hash and salted with the participant's username. This ensures that the same filename would not have the same hash between two different users and also that should a malicious user get hold of the SQLite database, they would not be able to decrypt the stored data easily. However, should they guess that the salt for the SHA-512 hash was the username they could use a process of trial and error to gain the list of images used. At no point during the login process is the hash stored in the database physically decoded. Rather the hashed image file names chosen by the user are compared with the ones in the database and if a complete match is found then authentication is counted as successful.

C9 has not been enabled on the system for the purposes of the experiment to see whether participants would be able to log in at all. It would however be trivial to include the ability

to block further attempts at logging in should a limit be reached on the number of failed logins attempted. The logistics of how the unlock procedure could be implemented would be partly the responsibility of the mobile operator. Companies already have systems in place to unlock phones where a PIN has been entered incorrectly a set number of times, so it would be an extension of this process. It would also be equally as easy to block for a set period of time to help prevent a guessing attack from succeeding.

C10 is satisfied in the application by not displaying on each page whether or not the participant has chosen the correct image from the page. Even when they complete the set, the user only knows whether or not the overall set of 4 image choices was correct or not; they are given no indication of which ones were correct or incorrect.

5.3.2 DrawASecret

The DrawASecret method as implemented uses the Canvas widget on android to create a background onto which the user can draw a line using their finger. Superimposed on this is a light grey grid which indicates the squares which will be recorded and stored by the method. Whenever the screen is refreshed, this grid is replaced on top of the basic canvas. A 5x5 grid is used here rather than the original 4x4 for a similar reason to Awase-E in that a 5x5 grid was felt to be usable given the larger size of the screen and would offer an obvious increase in security. A button was also provided to allow the user to reset the screen to start the drawing over if they had made a mistake. An example of this, along with a sample drawing is shown in Figure 26

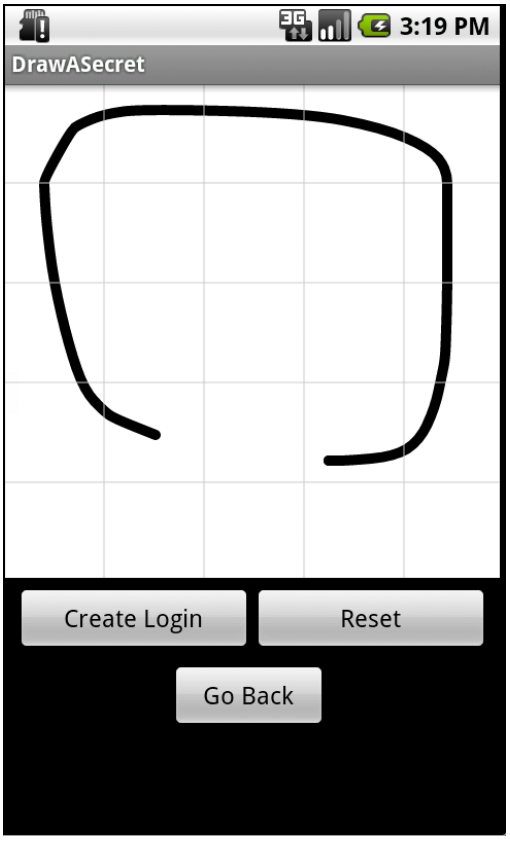


Figure 26: Create account stage of the DrawASecret method

Lines or points started by the user on one of the grid lines are discounted entirely and removed from the screen after drawing, although a warning message is displayed to the user to tell them what happened. This is to help prevent ambiguity in the storing of the grid path in the database, but also for the user when reproducing it both to confirm the password at the account creation stage and when logging on. A confirmation screen was added beyond the original specification to help protect against accidently brushing the screen causing an unintentional extra line on the password which requires the user to repeat the drawing before it is saved. The process diagrams for DrawASecret are shown in Figure 27 and Figure 28.

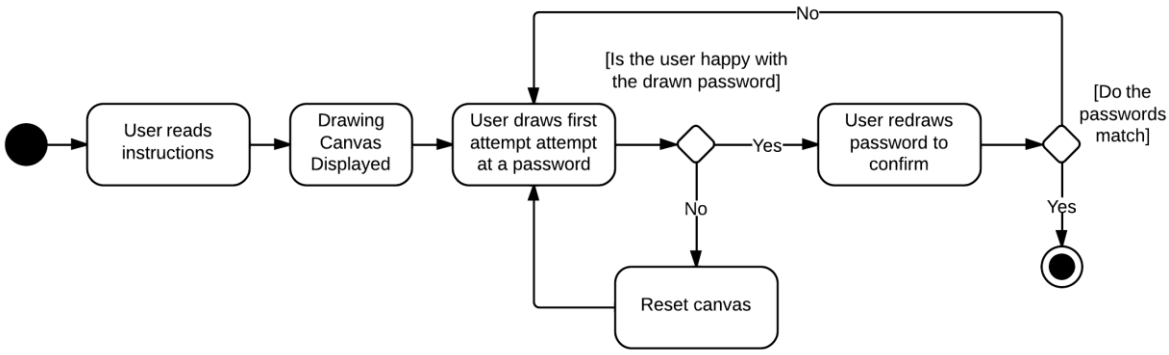


Figure 27: UML Activity Diagram for the Create Account phase of the DrawASecret method

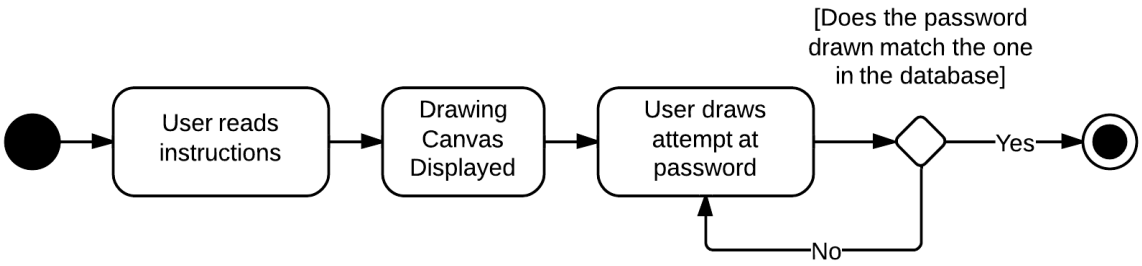


Figure 28: UML Activity Diagram for the Login phase of the DrawASecret method

Objective Criteria

Similar to Awase-E, the only options available to the user are to login using the method presented to them. There is also a check in place to ensure they use a sufficiently complex password to be secure. Currently this is set to a minimum of 4 data points in the database (so either two points or a line passing through three squares). This can be modified easily to increase the required complexity of the password. Any system error that occurs returns the use to the login screen rather than allowing entry through the application thus satisfying C7.

All of the squares stored on the SQLite database used for logging in are encrypted using the SHA-512 with the username salt as with Awase-E. However, there are only a limited number of possible hashes in any given DrawASecret password. These are simply the

number of squares available (25) combined with the start and stop markers, so 27. This would make a brute force attack to decode the data stored relatively easy to perform successfully.

A method was developed to ensure that in such a situation, the data decoded was not the actual grid numbers used for the password. For instance the a path going from grids numbers 1 to 2 to 3 could fairly easily be recognised as a line from the top left square moving two squares to the right. The method makes use of the same hashing method as before, but using the username. First the username is hashed using SHA-512 which produces a 128 character long string. This can be doubled by reversing the hash and concatenating it onto the end of the original to produce a 256 character long string. If blocks of three characters are taken from the final string then it is possible to repeat the overall string three times without a guaranteed repetition to produce a 768 character long string in the order given below and split as shown.

abc|ddc|baa|bcd|dcb|aab|cdd|cba

The characters produced from the hash are hexadecimal only so there are a total of only 4096 combinations to choose from (16^3). Assuming an entirely random spread there will always be some matches in this for each of the grids (for example, there is a greater than 99.9% chance of there being at least one identical set of characters in the grid) so when this happens, the particular grid reference is skipped in favour of the next block of 3 characters. This means that the maximum number of grids that can be used, assuming a square pattern, is 15 since 15^2 equal 225. Even this leaves room for up to 31 possible collisions which is so extremely unlikely it is safe to assume it will not happen (at about 21 collisions the probability is effectively zero, so by 31 it is safely unlikely). For a basic five square grid

there is a very small probability of failure. As with Awase-E, the blocking login for a certain number of failed login attempts has not been enabled.

Finally C10 is met by not showing the user which parts of their drawn password were correct and which were not. This did lead to some usability issues with people remembering the shapes of the password drawn, but not their position (detailed in Chapter Six).

5.3.3 PassPoints

The create account stage on PassPoints is split into two stages. First the user has to choose a background image for their account, before being taken on to the choosing points stage. The user creates an account using the gallery method in a similar manner to Awase-E, in that they have to scroll along to see all of the available images (Figure 29).

Figure 29: Selecting the background image for the PassPoints password

Once an image has been chosen, the user is forwarded to a different page on which they have to choose the areas on the screen to act as the password. This is implemented in a similar manner to DrawASecret using the canvas widget to allow the user to 'draw' onto the screen. This allows them to press on the screen to create a square which can be dragged to the correct area of the screen. Rather than storing the exact point at which the user has touched the screen, and calculating the area from that, the canvas itself was split into squares, and which ones of these squares were stored instead. An example of a user's set of PassPoints is shown in Figure 30.

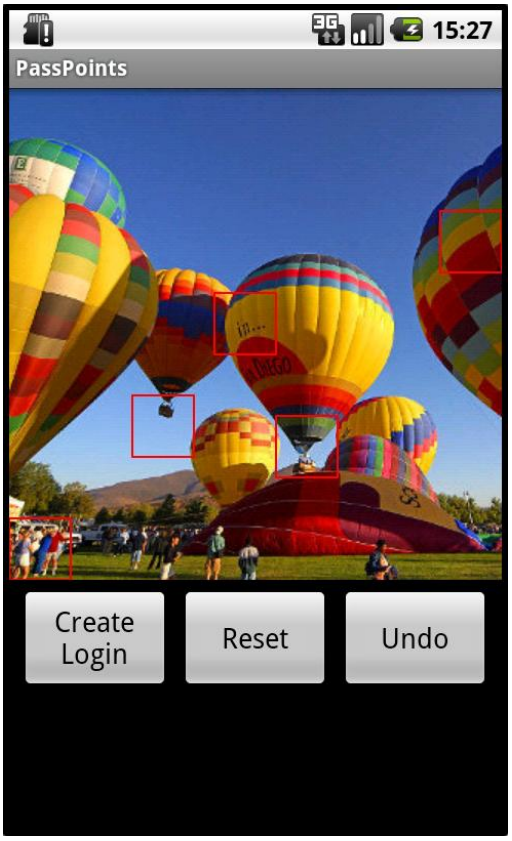


Figure 30: Creation of pass points

When selecting the areas, the user is in fact choosing grids on an unseen structure not included in the image displayed. To avoid the issues of a user choosing points too close to one or more of the gridlines (and so substantially changing the area required to confirm the grid position) a method of robust discretization (Birget et al., 2003) is used using three separate grids (see Figure 31). If the user selects a point on one of the gridlines, or within a certain unsafe distance from one, then the application automatically tries the next grid which is shifted in position from the first. By using three such grids, it is possibly to guarantee that any point on the screen is at least a safe distance from one of the grids.

The original implementation of this method did not use the robust discretization method, and instead relied on a simple circle created a fixed distance around the point chosen by the user. When confirming the password, the user simply had to press anywhere within this

circle to be authenticated. However this method relied on the data in the database either being "unhashed" or unencrypted, or a brute force style checking method whereby every point within the radius of the chosen point was checked and hashed to see if it matched the record stored in the database. The first of these options is unsecure and the second too computationally intensive for a mobile device, since even a 10 pixel radius could end up with over 300 hashes required for each point to be performed in order to confirm the user had picked a point within the required area. This was therefore discarded as a secure method of logging in.

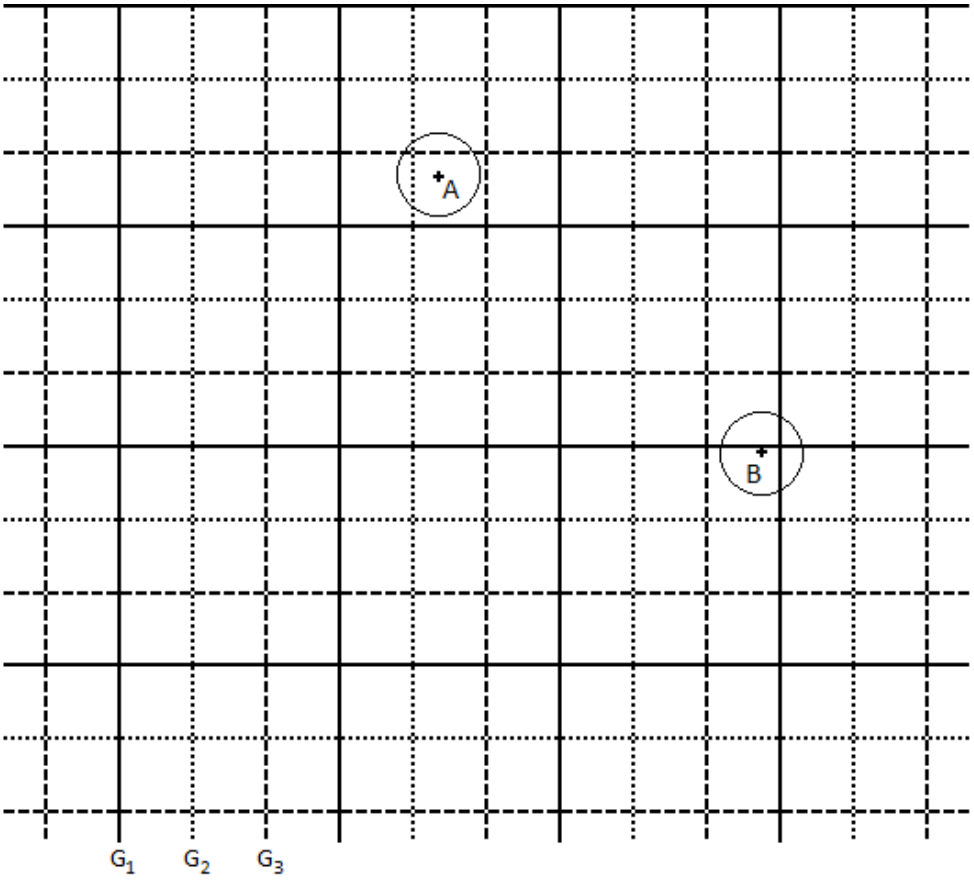


Figure 31: Robust discretization for choosing which grids points A and B will be assigned to

Figure 31 shows an example of how two different points would be assigned to different grids using robust discretization. In this example, G_1 is the top most grid, followed by G_2

and G_3 as the lower grids. Point A would be assigned to grid G_1 straight away since it is far enough away from the gridlines for G_1 to be 'safe', whereas point B is too close to grid G_1, so would be assigned to G_2.

The screen was split into three sets of grids, the 'top' grid being 8 squares across, and the other two being 9 (to accommodate the overlap at the edges). This appeared to be the best compromise between a more secure version with more squares and consequent lack of usability when the squares are too small to choose accurately. Smaller squares also lead to issues with the safe distance away from a gridline required for a different grid to be used. In order to space the grids apart evenly, the safe distance tends towards to 1 pixel almost as the number of grids are increased, making it nearly impossible to accurately choose the same grid each time. However, this set of 8 squares was shown in the experiment to be too small an area for the majority of people to cope with easily, so would probably have to be decreased (enlarging the squares) in any future implementation.

Using the implementation in Experiment 1, there were total of 226 available squares which could be chosen by the user. Even allowing for the fact that no overlapping squares were permitted to be chosen, the number of combinations of squares for set of 5 PassPoints is of the order of 10^{11} (more than 3.85×10^{11}), clearly higher than that of AwaseE

When confirming the account during the account creation phase, it was found during testing that it was extremely difficult to accurately recreate the exact grid chosen on the first attempt. To rectify this, the application was altered so that whilst confirming, a user only has to choose a point within the original grid chosen first. The application recognises this is one of the user's original points and automatically moves the grid into this correct

position. This led to an issue with overlapping grids as it is impossible to say which of the grids a user meant if there are two overlapping at a given point. To prevent this from being an issue, and also to ensure that the grid points chosen by the user were distinct, the ability to choose overlapping grids on the image was removed altogether.

When logging in on PassPoints, the user has to choose the same background image they chose when creating their account. This helps to prevent the issue of people always choosing the same the same points on their image. An attacker would therefore also have to guess which image the user chose as well as the points, although this is would only be from a much reduced set of possible images than the points themselves. The images displayed as choices for the background, as with those for the Awase-E method, will always be randomised when presented to the user.

Once the user has chosen a point, the best grid position is displayed as one of the areas they have to remember on the screen as was shown in Figure 30. This method may however influence people to choose similar areas on the screen. For instance, whilst testing the application, an image of a cat was used, and all of the people who picked this as the background image chose the same five points on the cat's face, both ears, both eyes and the nose. As such, images have to be chosen carefully to try not guide the user into always choosing the same points, or at least having a large enough set of obvious points for it to not be easy to guess which will be chosen. The process diagrams for this method are shown in Figure 32 and Figure 33.

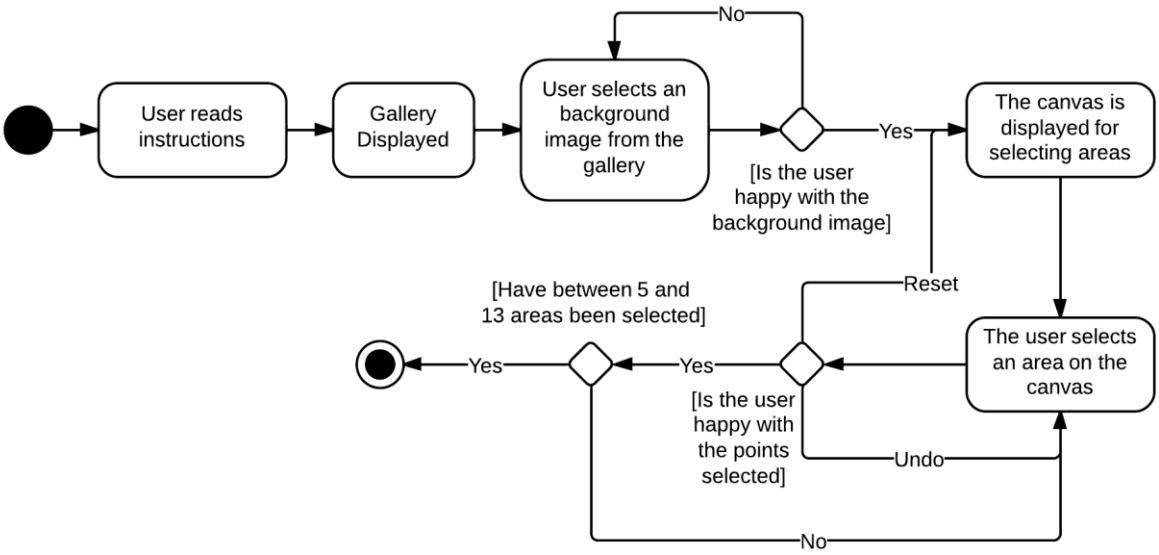


Figure 32: UML Activity Diagram for the Create Account phase of the PassPoints method

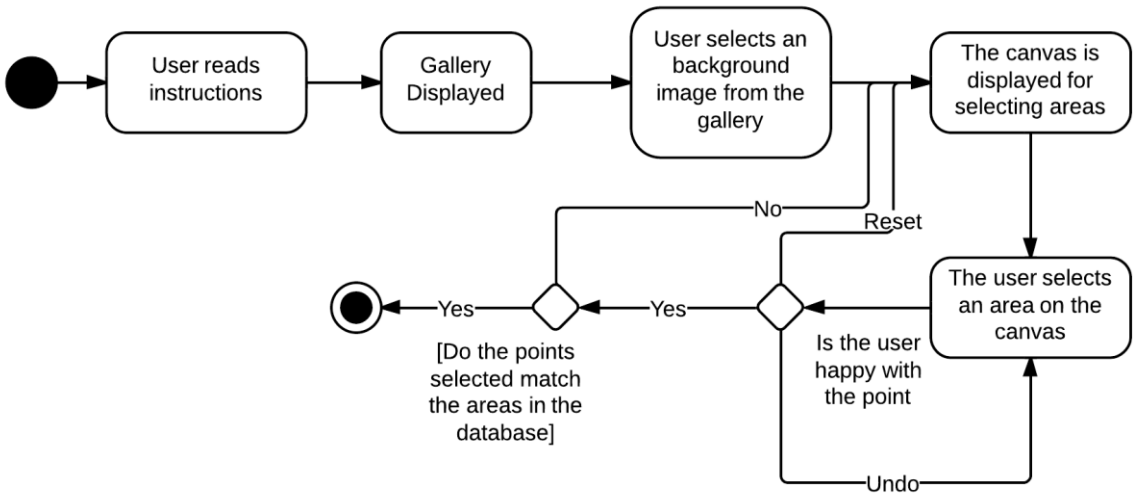


Figure 33: UML Activity Diagram for the Login phase of the PassPoints method

When logging in, the user has to choose these areas from on the background image in order to be authenticated. Anywhere within the initial square chosen by the user is accepted as correct for that point. An example of what the user would see after having selected the five correct points is given in Figure 34.

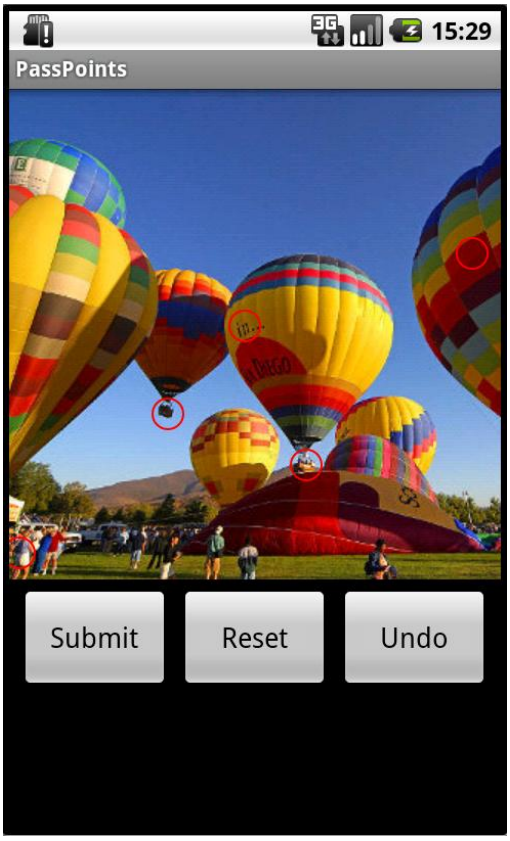


Figure 34: Logging in using the PassPoints method

The PassPoints can be confirmed in any order, rather than specifically in the order they were chosen, although obviously for an additional level of security, requiring the order to be remembered could be implemented into the method.

Objective Criteria

As with Awase-E and DrawASecret, C1 is satisfied within the implementation by only having options available to the user which will allow for the most secure login. A check is in place to ensure the set of PassPoints is complex enough by ensuring that people choose at least five PassPoints.

Again, as with DrawASecret and DrawASecret, any system error automatically returns the user to the beginning login screen rather than granting access, satisfying C7.

For C8, PassPoints uses the same method as DrawASecret to randomise the allocation of strings to represent each of the underlying grid numbers. However since there are more numbers possible there is a slightly higher chance that the process would run out of strings before being able to reference each grid (although even with a fairly small nine square grid it is still near zero). A mobile screen is unlikely to ever reach a point where 15 grids can be accurately drawn on so this would not have caused any problem in this experiment. But, if the application was ported to a device with a larger screen it would start becoming an issue if 15 or more grids were required. Reducing the number of grids to 14 substantially decreases the chances of running out of random three character letters generated by the hashing method. Even at this point the grids would be far too small to be reasonably usable on a standard mobile screen so these can safely be ignored.

C9 has not been enabled currently on this system.

C10 is satisfied in the application by not displaying whether or not the user has chosen the correct background image or PassPoints. However if they choose the wrong background image they would never be able to log in, even if they were to somehow choose the correct points. In either case, which PassPoints the user has chosen are also not confirmed as correct or false, only whether the complete set was. As with DrawASecret this led in the experiment to some people remembering the generic pattern of PassPoints well enough but not specifically enough to put the points into the correct positions (e.g. whether or not a chosen point was to the left or the right of an object in the picture).

5.3.4 Implementation details common to all methods for logging actions

For the purposes of the experiment a lot of data was logged for later analysis, particularly on the methods using a canvas (DrawASecret and PassPoints). These methods required logging of the movements made by the user on the screen itself, which for a line can result in a substantial amount of data being stored within the database. Writing this to the database caused a delay, so a display timer had to be added to prevent participants from attempting to press options on the screen multiple times. These writes are all done when the user submits an attempt at creating an account or login, be it successful or not to stop it happening continually whilst the user is attempting to use the program.

All of the images used, both as the background images for PassPoints and the individual pass images for Awase-E, were collected from one of Wikimedia commons, PDPhoto.org or FreeDigitalPhotos.net. All of the images are public domain images freely available to be used for any purposes.

The following chapter presents the results and analysis of the experiment detailed in this chapter.

6 Chapter Six – Experiment 1 - Results and Discussion

This chapter presents and discusses the results from Experiment 1 which was described in Chapter Five. Firstly, the questionnaire responses (Dataset 1) are presented and analysed according to the usability criteria to which they relate. Datasets 2 and 3 are also analysed to determine whether or not they support the hypotheses presented in Section 5.2.1. Other factors that might have influenced the results of the experiment are investigated, such as psychological influences on the participants, along with the known limitations of an experiment of this type. The criteria themselves are also re-examined to determine whether they were sufficient to evaluate whether the methods were usable.

6.1 Method Comparison – Dataset 1

The two groups of participants' data from Dataset 1 were considered separately due to the expected large difference in technical skill between the people participating within each group. Within this, each of the results for each criterion was analysed to see if there was any statistically significant difference between the methods.

Each of the questions was scored using a Likert scale from 1 to 7 as the answer. Once all of the questionnaires had been completed, a score was generated for each criterion for each of the methods based on the questionnaire. Where more than one question related to a criterion an average was taken in order to produce the score. Questions 9 to 11 and 12 to 13 were regarded as two separate usability variables, and scores generated for these separately.

The scores from the questionnaire were then analysed using the method taken from the work done on Cross-Over studies by Senn (2002). The experiment was designed in such a way to allow this and so allow for all three methods to be compared directly against one another, which is also the reason a standard t-test could not be used. This type of analysis is particularly suited to experiments where there are three or more treatments or methods being evaluated at once, and the goal is to confirm the null hypothesis that there is no difference between them without having to perform a pairwise comparison of each of the treatments. Using Senn's method, an F-test statistic, or F ratio, is produced for the experiment as a whole, and this can be compared to statistical tables for the appropriate number of degrees of freedom and accuracy to indicate whether or not there is a statistically significant difference across the results from the entire experiment. If the calculated F ratio is higher than that found in the statistical table, then there can be said to be a statistically significant difference between the results for the different methods. In this case, the calculation will be performed multiple times, both due to the two groups of participants being separated, and within each group, once for each of the different criteria being analysed.

The C2 criterion had to be analysed slightly differently due to the nature of the results gathered in. This criterion related directly to question 1 on the questionnaire asking the participant how they dealt with errors made during the process of creating their account and logging in. Some of the participants in each of the groups did not make any errors whilst using the methods, and so either picked a middle range number or answered "not applicable" for this question. As such, the averages taken for this criterion are only calculated over the number of participants who were able to give a legitimate answer. This also affected the number of degrees of freedom and so the critical F ratios for this criterion

when compared to the others. Some participants didn't use the options to undo or reset specifically, but did give a score based on how useful they thought the abilities would be, so for these participants the score given was kept.

From Group A, three of the C2 scores for Awase-E were removed in this manner, two of the DrawASecret and none from PassPoints. Only one participant's scores for C2 on the Awase-E method had to be removed from Group B, and none for the other two methods.

F-Test process

The results for criterion C3 from Group A is used as an example here to demonstrate the method used to calculate the F ratio for each of the criteria for each group. C3 is used as the example criterion here, rather than C2, due to the additional complexities of having to balance the number of users for the C2 criterion. Table 8 shows the usability scores given from the questionnaire for the C3 criterion for each of the participants in Group A, and a corresponding table for Group B is presented as Table 9 for completeness.

Participant	Method			
	Order	Awase-E	DrawASecret	PassPoints
S1	ADP	7	7	7
S2	APD	7	3.5	7
S3	DAP	6	6	6
S4	DPA	6	6	6
S5	PAD	6	5	5.5
S6	PDA	7	6	7
S7	ADP	7	5.5	6

S8	APD	6.5	7	7
S9	DAP	5.5	6.5	4.5
S10	DPA	6	6.5	7
S11	PAD	6	6	6
S12	PDA	3	7	6
S13	ADP	6.5	7	6.5
S14	APD	7	7	2
S15	DAP	7	7	7
S16	DPA	6	6	6
S17	PAD	7	7	6.5
S18	PDA	7	7	7

Table 8: The scores for criterion C3 for each of the authentication methods from the participants in Group A

Participant	Method			
	Order	Awase-E	DrawASecret	PassPoints
GP1	ADP	6.5	6.5	6.5
GP2	APD	7	5	5.5
GP3	DAP	6	5	5
GP4	DPA	6.5	5.5	6
GP5	PAD	6	4	6
GP6	PDA	6	6	6
GP7	ADP	7	7	7
GP8	APD	5.5	6.5	5.5
GP9	DAP	5.5	5	3

GP10	DPA	7	7	7
GP11	PAD	7	6	5.5
GP12	PDA	7	7	4.5
GP13	ADP	7	7	7
GP14	APD	6.5	7	7
GP15	DAP	7	7	7
GP16	DPA	6	7	7
GP17	PAD	6.5	6	7
GP18	PDA	4	1	4

Table 9: The scores for criterion C3 for each of the authentication methods from the participants in Group B

From Senn (2002), if there are n participants, and k methods, then with the overall mean as $\bar{Y}_{..}$, the mean response over all participants for the i-th method is $\bar{Y}_{i.}$. and the mean over all methods for the j-th participant is $\bar{Y}_{.j}$, then formulae for the sums of squares are

SSMethods = $n\sum(\bar{Y}_{i.} - \bar{Y}_{..})^2$

SSParticipants = $k\sum(\bar{Y}_{.j} - \bar{Y}_{..})^2$

SSError = $\sum(\bar{Y}_{ij} - \bar{Y}_{.j} - \bar{Y}_{i.} + \bar{Y}_{..})^2$

SSTotal = $\sum\sum(\bar{Y}_{ij} - \bar{Y}_{..})^2$

Specifically for C3 here, this means that the overall mean for all methods for C3 comes out as 6.23. The totals for each of the methods are averaged and this average subtracted from the overall mean. Squaring each new value and summing them gives the Sum Squared value for methods as 0.40. Using a similar method for each of the participants gives that value as 18.02. The total sum of squares is found by subtracting the overall mean from each of the values in the table, squaring and summing, and comes up with an answer of 56.86.

This gives the sum squared error for the whole experiment as 38.444 when using the formula:

SSTotal = SSMethods + SSParticipants + SSError

Source	Sum of Squares	Mean Square			F Ratio
		Freedom		Score	
Participant	18.02	n-1=	17	1.06	0.93
Model	0.40	k-1=	2	0.20	0.18
Error	38.44	(n-1)(k-1)=	34	1.13	
Total	56.86	nk-1=	53	1.07	

Table 10: F Ratio calculations for criterion C3 from Group A

As is shown in Table 10, these values can be used to calculate the F Ratio for the whole experiment in terms of both the participants and the methods. First the degrees of freedom are taken into account for each of the calculated sum squares to give a mean squared value for each. There are 18 participants, and so 17 degrees of freedom gives the mean squared value for the participants as 1.06. Three methods were evaluated, so 2 degrees of freedom gives 0.20 for the mean squared method effect. The mean squared error takes into account both methods and participants and so the degrees of freedom are multiplied to give 34 and so a mean squared error of 1.13. The F Ratios can now be calculated by dividing the mean squared effect by the mean squared error. This gives 0.94 for the participants and 0.18 for the methods. The F Ratio for participants is of little importance since it is assumed the participants will vary anyway.

The method F Ratio is more significant since it shows that there is no statistically significant difference between any of the methods in terms of the usability scores awarded. A high enough F Ratio would imply this is not true and so there is a difference between the

methods. However the critical F ratio for 17 and 2 degrees of freedom is 3.28 and so at least in terms of C3 for Group A, there is no difference between any of the methods. A full table of F Ratio ratios for Group A can be found in Table 11 and for Group B in Table 12.

Criteria	Mean Score			Method F Ratio	Statistically Significant
	Awase-E	DrawASecret	PassPoints		
C2	5.80	5.87	5.6	0.26	No (Critical F: 3.34)
C3	6.31	6.28	6.11	0.18	No
C4	6.44	6.56	6.11	0.96	No
C5	6.25	5.72	4.83	9.22	Yes
C6	6.28	6.06	5.72	1.41	No
C11	6	5.89	4.83	7.46	Yes
Gen1	5.78	5.24	4.09	13.34	Yes
Gen2	6.19	5.81	4.92	7.72	Yes

Table 11: Mean scores and statistical significance test for all of the criteria for the Group A

Criteria	Mean Score			Method F Ratio	Statistically Significant
	Awase-E	DrawASecret	PassPoints		
C2	5.44	5	5.17	0.51	No (Critical F: 3.29)
C3	6.33	5.86	5.92	2.00	No
C4	6.22	6.64	5.94	1.88	No
C5	5.89	4.78	4.44	4.6	Yes
C6	6.44	5.39	5.56	3.19	No
C11	6	5.5	5.22	1.81	No

Gen1	5.09	4.11	4.33	2.42	No
Gen2	5.89	4.31	4.58	8.94	Yes

Table 12: Mean scores and statistical significance test for all of the criteria for the Group B

For all of the criteria except C2, this critical F ratio is 3.28 (for 17 and 2 degrees of freedom), and is shown in Tables 11 and 12 for C2, for the appropriate degrees of freedom for each group. If the F ratio from the experiment is calculated to be above the critical F ratio, then it would imply there is a statistically significant difference between the methods for that specific criterion. For Group A, this is the case for the C5, C11, Gen1 and Gen2 criteria, and for Group B, only the C5 and Gen2 criteria. In order to ascertain which of the methods were considered more usable than the others for these criteria, a pairwise comparison of each of them would need to be carried out to see which of the individual methods can be shown to be significantly different.

Looking more deeply into the comparisons between specific methods a basic estimators approach can be used to show the difference between pairs of methods. As an example of the process taken, the comparison of Awase-E vs. PassPoints is examined on the C5 criterion since, unlike C3, there was shown to be a statistically significant variance across all three methods.

The difference between the scores for Awase-E and PassPoints is taken for each of the participants and this is averaged for each of the orders the participants used (i.e. ADP, APD etc). This gives a better estimation of the overall difference whilst still allowing for any variation that the effect period may have on the results. This result of this is shown in Table 13 (again for the Awase-E/PassPoints comparison on C5).

Using this method, the period effects between the different orders cancel out over the entire result set allowing the actual differences between methods to be analysed rather than having them influenced by the order in which the methods were evaluated.

Sequence	Number of Participants	Degrees of Freedom	Mean	Variance	Corrected Sum of Squares
ADP	3	2	1.83	0.58	1.17
APD	3	2	3	7	14
DAP	3	2	0.5	0.25	0.5
DPA	3	2	1.3	3.08	6.17
PAD	3	2	0.83	2.33	4.67
PDA	3	2	1	0.25	0.5
Total	18	12	1.42	13.5	27

Table 13: An example comparison between Awase-E and PassPoints showing the variance and corrected sum of squares for the Group A for each of the orders participants were tested in looking specifically at criterion C5

The variance of each of the sequences can also be calculated allowing the total corrected sum of squares to be calculated from the corrected sum of squares for each sequence. This sum of the correct sum of squares is not influenced by any period effects and allows a basic estimation of the variance of a basic estimator by dividing it by the total number of degrees of freedom, in this case 12 giving 2.25.

Using the formula variance $\sigma^2 = \sum_s (\frac{1}{m_s})/r^2$ allows the variance of the method effect to be calculated as 0.125, the square root of which gives the standard error of 0.35. To calculate the confidence limits, a t value of 2.18 is used, based on the statistical tables for

an F distribution corresponding to 12 degrees of freedom and a 2.5% tail area. Here the degrees of freedom are calculated by taking the sum of the degrees of freedom for each permutation of the different orders the methods were tested in. Since six sets were used, each with three participants, the number of degrees of freedom for each set is two, totalling to 12. The t value is multiplied by the standard error of 0.35 to give 0.76 which can be added and subtracted from the point estimate of the mean to give confidence limits for the method effect (τ) of $0.65 \leq \tau \leq 2.19$. This means there is a 95% chance that the method effect between Awase-E and PassPoints showed an improvement of between 0.65 and 2.19 on Awase-E over PassPoints.

These confidence intervals can be calculated for each of the comparisons of methods where the F-test shows there is a significant variation in the results over the entire experiment. Where the F-test does not show a significant variation, the confidence intervals cannot be said to be accurate enough to show a general difference between the methods, when considering their performance against the specific criterion in question.

Table 14 shows a comparison of the individual methods for each of the criteria where there was shown to be a statistically significant difference across all three methods. For the purpose of the method differences, the difference between methods is taken in the order the comparison is made in the Method Comparison column.

Criteria	Group	Method Comparison	Method Difference	σ^2	SE	Limits	Statistically Significant
C11	A	A vs. D	0.11	2.39	0.36	$-0.63 \leq \tau \leq 0.85$	No

		A vs. P	0.17	2.11	0.34	$0.42 \leq \tau \leq 1.91$	Yes
		D vs. P	1.06	2.94	0.40	$0.17 \leq \tau \leq 1.94$	Yes
Gen1	A	A vs. D	0.54	2.82	0.40	$-0.33 \leq \tau \leq 1.40$	No
		A vs. P	0.69	4.22	0.48	$0.63 \leq \tau \leq 2.74$	Yes
		D vs. P	0.15	7.55	0.65	$-0.26 \leq \tau \leq 2.56$	No
Gen2	B	A vs. D	1.58	4.18	0.48	$0.53 \leq \tau \leq 2.63$	Yes
		A vs. P	1.31	1.36	0.28	$0.71 \leq \tau \leq 1.91$	Yes
		D vs. P	-0.28	2.99	0.41	$-0.17 \leq \tau \leq 0.61$	No
Gen2	A	A vs. D	0.39	1.44	0.28	$-0.22 \leq \tau \leq 1.00$	No
		A vs. P	1.28	2.46	0.37	$0.47 \leq \tau \leq 2.08$	Yes
		D vs. P	0.89	5.08	0.53	$-0.27 \leq \tau \leq 2.05$	No

Table 14: Comparison of the individual methods using correct sum of squares for the six criteria shown to be statistically significantly different across all three methods

A pairwise comparison of each of the methods for the criteria, other than C5, which were found to be statistically significantly different, is shown in Table 14. This shows a pairwise comparison between the methods. Confidence limits are calculated in order to show how much of a difference was indicated when comparing the methods against each other. Where the confidence limits for a pairwise comparison ranges from a negative to a positive number, the result cannot be statistically significantly different since it is not clear whether there was a guaranteed difference in scores between the two methods. This leaves the following statistically significant differences between methods based on the criteria, where such a difference implies that the application with the higher mean is more usable than the other:

- C6 Group A: Both Awase-E and DrawASecret are more usable than PassPoints, but there is no difference between Awase-E and DrawASecret.
- Gen1 Group A: Awase-E is more usable than PassPoints.
- Gen2 Group B: Awase-E is more usable than both DrawASecret and PassPoints but there is no difference between DrawASecret and PassPoints.
- Gen2 Group A: Awase-E is more usable than PassPoints.

Overall, this shows the null hypothesis can only be fully rejected for the Gen2 and C5 criteria, as well as criteria C6 and Gen1 in Group A on the basis of the usability criteria. For all of the other criteria, there was no significant difference across the three methods and so H01 is true for these. The qualitative data (see Section 6.4) does support H01 being false however.

6.1.1 Comparison of Groups A and B

In order to compare the two groups directly against each other, the scores for each criterion for each method were compared using an unpaired two-tailed t-test. The purpose of this test is to determine whether one group scored the methods higher than the other group, which would indicate which of the groups found the methods easier. This relates directly to H02 which is concerned with the relative difference in technological ability and experience between the two groups.

Method	Criteria	Group A Average	Group B Average	Difference (Group A – Group B	T Statistic
Awase-E	C2	5.8	5.53	0.06	0.62
	C3	6.31	6.33	-0.03	0.93
	C4	6.44	6.22	0.22	0.45
	C5	6.24	5.89	0.36	0.32
	C6	6.28	6.44	-0.17	0.65
	C11	6	6	0	1
	Gen1	5.78	5.09	0.69	0.17
	Gen2	6.19	5.89	0.31	0.39
DrawASecret	C2	5.94	5	0.72	0.17
	C3	6.28	5.86	0.42	0.33
	C4	6.56	6.64	-0.08	0.64
	C5	5.72	4.78	0.94	0.1
	C6	6.06	5.39	0.67	0.23
	C11	5.89	5.5	0.39	0.45

	Gen1	5.24	4.11	1.13	0.06
	Gen2	5.81	4.31	1.5	0.01
PassPoints	C2	5.72	5.17	0.56	0.27
	C3	6.11	5.92	0.19	0.63
	C4	6.11	5.94	0.17	0.72
	C5	4.83	4.44	0.36	0.5
	C6	5.72	5.56	0.17	0.77
	C11	4.83	5.22	-0.39	0.52
	Gen1	4.09	4.33	-0.24	0.66
	Gen2	4.92	4.58	0.33	0.57

Table 15: T-test comparison of the usability scores from each of the two groups

Table 15 shows the difference in the average usability scores for each criterion from each of the two groups, along with the t-test comparison of the two sets of data. Only one of the criteria, Gen2 for the DrawASecret method, has a t statistic below 0.05 so only this can be said to be a statistically significant difference. However, with a large number of tests being performed, it is likely that one of the results will be shown to be statistically significant by chance, rather than due to an overall trend. With so many of the criteria showing no statistically significant difference, the null hypothesis H02 cannot be rejected.

There are indications that participants in Group A found the methods more usable than those in Group B. When looking at the average scores for each group in Table 15, only five of the 24 comparisons showed Group B scoring higher than Group A. Although the scores were not different enough to be statistically significant, the high percentage of criteria with a higher average from Group A indicates this may be an overall trend.

6.1.2 Discussion

Over the whole experiment, most of the average scores are reasonably high, with no single criterion having an average less than four. Of the original criteria which were not excluded, only C11 for Group A showed a statistically significant difference across the entire experiment.

Looking at the criteria individually, scores for C2 varied very little across both participant groups, indicating there was very little different between the three methods in terms of this criterion. Both groups tended to score highly with those from Group A scoring marginally higher than those from Group B on average, meaning for the most part the purpose of the application was clear as was how to go about using it.

Similarly for both C3 and C4, the average scores were generally high (>5 in all cases) with no statistically significant difference between the methods. This shows that the participants felt the information displayed to them was easy to understand and not overly complicated or containing unimportant information as per C3, and colours within the applications did not distract from the main purpose.

C5 for both groups did show a significant difference between the three applications. Awase-E received the highest score for this criterion on average, followed by DrawASecret and finally PassPoints. For this reason it is excluded from the method comparisons shown in Table 14. C5 related to the speed of the application so one of the probable causes for this difference would be the logging enabled on the device. To record the choices and drawings made by the users, all movements on the touch screen and

buttons pressed were logged into a SQLite database on the device. With DrawASecret and PassPoints, this involved logging the individual movements made on the screen in order to accurately record drawings or motions made by the participant's finger. For even a relatively simple line this would involve a significant amount of data being added into the database, which would not apply for Awase-E where only button presses needed to be recorded. As such, the amount of time taken to add the data into each of the applications would be drastically smaller for Awase-E and so would require less waiting on the participant's part whilst it records it.

C6 showed very little variation within Group A, with the average scores near or above 6 for all three applications and so no statistically significant difference. There was a wider range of scores given by participants from Group B with Awase-E scoring better than the other two methods. However the F Ratio although close, is not high enough to be able to say this is a statistically significant difference to within 5%. This could be a slight indication that users in Group B found Awase-E easier to cope with mentally than the other two methods, which would fit with the idea that methods involving recognition (Awase-E) would perform better than those involving recall only (DrawASecret and PassPoints).

C11 returned different results between the two groups of users. Group B scored the criteria quite highly (always greater than five) with no significant difference between methods. Participants in Group A, however, scored Awase-E and DrawASecret significantly higher than the PassPoints application. C11 concerned the methods' effectiveness on a small screen and was primarily approached through question 8 on the questionnaire. Those participants in Group A were generally of a younger generation than those people in Group B, and would have been more familiar with mobile phone or smartphone technology. Thus

they may have been in a better position to decide if the small screen was a hindrance or not. The older people within the study generally felt that the screen was the largest available on a mobile device and so may have scored this question higher on those grounds. In addition, those in Group A might be more used to the larger screens of a desktop or laptop computer, which participants in Group B, although they may have used them, are unlikely to have been exposed to as much time using a computer than those of a younger generation.

As with C11, the variation in the Gen1 results was larger in Group A across the three applications than Group B, again possibly due to the wider experience students have with mobile phones in general, allowing them to make a more informed decision. However it may be that participants in Group A found using the Awase-E method easier than the other two. This could be caused by a difference in the ability for younger people to recall rather than to recognise images, for instance being better at recognition over recall, whereas older people often have worse memories and be less at ease with the application. Users will always like an application more if they feel they are capable of actually using it effectively, this being the user satisfaction aspect of usability as a whole.

It could also be that younger people are more opinionated about what they do and do not like in terms of applications on phones and are thus are willing to give lower scores to things they do not like. Older people with less experience might not be able to distinguish the benefit of one type over the other simply through lack of other experience, and would therefore attribute a lack of usability within the application to their own shortcomings.

For Gen2, both groups showed a significant variation in the scores across the three different applications with both cases showing Awase-E as scoring the best. For the Group A, this was followed by DrawASecret and then PassPoints whilst for Group B received the second highest scored followed by DrawASecret. This fits in with many of the comments made by the participants during the course of the experiment saying they did not like PassPoints at all, or found it confusing, as well as being unable to accurately use DrawASecret. Hence many of them generally preferred Awase-E as their favourite login method of the three.

6.2 Order Effect – Dataset 1

The order in which the evaluations were undertaken can be shown to have an effect on the usability and memorability of each of the methods. Psychological studies have shown that tests undertaken in the middle of a set, in this case of three, score lower than those taken first or last (Deese & Kaufman, 1957). This is due to the first test to be taken making the largest impression in the user's mind and the final one being the 'freshest' simply because it is the most recent. Table 16 shows the average scores for each of the methods across all of the criteria in each of the testing positions.

Position	Group A			Group B		
	Awase-E	DrawASecret	PassPoints	Awase-E	DrawASecret	PassPoints
1	6.34	5.59	5.77	6.18	5	4.39
2	5.86	6.40	4.54	5.58	5.58	5.48
3	6.08	5.74	5.57	5.98	5.01	5.57

Table 16: Average score across all criteria for methods in different positions in the testing order

Although some of the methods do fit into the pattern predicted by the theory, specifically Awase-E for Group B and PassPoints for Group A, there is no generalised trend across all of the applications. In fact in some cases the results directly contradict the theory; DrawASecret for Group A scored higher in the middle position than when it was in the first or third position.

The reason for this may be that the effect of order on the experiment is too subtle to be picked up in the results. The original experiments into this effect were done with people memorising blocks of text or words with which for the most part the participants would have been familiar with the concept (Deese & Kaufman, 1957). This experiment involved a varying learning curve for each participant meaning the effect the order may have had on their ability to remember the password would be drowned out by the often negative effect of being unfamiliar with the system and concepts involved.

There was no consistency between the two groups of people in terms of which methods showed the effect. The Awase-E results came closest to showing some sort of order effect with the first and third experiments performing better than the second one for both groups, however the differences between the method in each position was not always statistically significant, as shown in tables 17 and 18 .

Comparison of positions	Awase-E		DrawASecret		PassPoints	
	Difference	P Value	Difference	P Value	Difference	P Value
1 vs. 2	0.48	0.082	-0.82	0.009	1.23	0.017
3 vs. 2	0.23	0.259	-0.67	0.001	1.03	0.039
1 vs. 3	0.26	0.259	-0.15	0.567	0.20	0.568

Table 17: T-test comparison of the difference between each method in each position separated between blocks of participants for Group A.

Comparison of positions	Awase-E		DrawASecret		PassPoints	
	Difference	P Value	Difference	P Value	Difference	P Value
1 vs. 2	0.60	0.092	-0.58	0.23	-1.09	0.029
3 vs. 2	0.40	0.247	-0.57	0.176	0.09	0.793
1 vs. 3	0.40	0.372	-0.01	0.979	-1.17	0.014

Table 18: T-test comparison of the difference between each method in each position separated between blocks of participants for the Group B

Tables 17 and 18 show the differences between the methods when comparing the positions each method evaluated and the P Value calculated from an unpaired t-test between the different sets of results. None of the methods show a universal difference in different positions across groups as would be represented by a P value lower than 0.05 for each comparison. However, some individual comparisons can be drawn.

From the theory, it would be expected that the methods would perform significantly better in position 1 than in position 2 and significantly better in position 3 than position 2. This ought to be represented by a positive value for the difference column for both 1 vs. 2 and 3 vs. 2, and a low P value for each. The difference between position 1 and position 3 would most likely not be significant to the results.

Awase-E in both groups did show a significant difference between scores when testing in positions 1 and 2. However, although the average scores were better in position 3 when compared to position 2, this difference is not significant.

DrawASecret went against the expected results by being given higher scores in position 2 than in either position 1 or 3, for both groups. Furthermore, this was a significant difference for both comparisons in Group A.

PassPoints for Group A does support the original theory with positive significant differences between positions 1 and 2 and 3 and 2. However, for Group B both positions 2 and 3 scored significantly better than position 1 with no significant difference between positions 3 and 2. This may be due to the relative difficulty people experienced when using PassPoints. Where this was their first experience of using a touchscreen phone, this might have given them a larger shock, or higher learning curve, immediately, leading to the lower usability scores given.

Interestingly, the comparison between positions 1 and 2 for Group A is always a positive significant difference indicating that when tested in position 1, the methods performed better than in position 2. However, there is no consistency in the comparison between positions 3 and 2. Although this does show partial support for the theory, the lack of consistency would indicate that there may be another reason for the results beyond merely that of the order.

Overall there is little evidence to support the order effect being present within the results found from the experiment making the null hypothesis H03 true, and HA3 false. Although some of the results do seem to show the order effect taking place, there is no consistency across the experiment to be able to say it was present. This is likely due to the overall difficulty most participants had with all of the methods. The Order Effect is a subtle difference in the results of several methods or methods being compared in different orders, and it is likely that this effect is being drowned out by other factors within the experiment, such as subject's unfamiliarity with the technology.

6.3 Logging Analysis – Dataset 2

Dataset two was collected from the logged data on the mobile device detailing the actions taken by each participant during the experiment. This shows the time taken for each stage of the treatment, along with the number of errors made by the participant both when creating their account and logging in.

	Create Account	Login				
	Average Time (s)	Successful Logins	Average Time (s)		Errors	
			All	Successful	All	Successful
Group A	57.9	17	41.6	36.2	1.2	0.8
Group B	92.3	16	85.9	64.8	1.7	1.1

Table 19: Average times for creating an account and logging in for participants using the Awase-E method

	Create Account		Login				
	Average Time (s)	Errors	Successful Logins	Average Time (s)		Errors	
				All	Successful	All	Successful
Group A	77.1	1.4	16	13.6	7.6	0.8	0.2
Group B	134.6	1.8	14	31.2	9.7	1.8	0.4

Table 20: Average times for creating an account and logging in for participants using the DrawASecret method

	Create Account		Login				
	Average Time (s)	Errors	Successful Logins	Average Time(s)		Errors	
				All	Successful	All	Successful
Group A	129.9	1	15	52.6	46.7	2.6	1.3
Group B	177.7	1.3	11	80.1	66.6	2.7	0.8

Table 21: Average times for creating an account and logging in for participants using the PassPoints method

Tables 19 to 21 show the average times taken by the participants to perform each part of the experiment for each group. Also shown is the average number of errors made by the participants whilst logging in. Separate averages are shown for the set of participants who were able to log in successfully using the application. This was done because the number of errors for length of time taken was significantly higher for those participants who were unable to log in at all.

6.3.1 Number of Errors

As was expected, the Group A was able to complete the tasks on all three applications faster than Group B and with fewer errors. The averages are given for both the overall group and separately for only the participants that were able to log in successfully. Participants who were unable to log in first time often made several subsequent attempts and so would skew the results for the group as a whole. For example, one of the participants had 21 attempts using the PassPoints method before giving up, spending more than double the average time for the other participants on logging in using this method. This is reflected in the large difference in the average time for each login when comparing the successful group to the whole for each of the applications, compared to little or no difference for the creating account phase. This shows that the difference in time is more due to the inability to remember the password rather than an inability to use the system itself.

Those people who were able to log in generally were able to do so on their first or second attempt on each of the methods evaluated. This will be due to them being able to remember their password effectively. However, these tables do not show any particular

indication of how complicated each password was. For the DrawASecret method, there did not appear to be any direct relationship between the complexity of the password and whether or not the participant was able to log in. Although the average password complexity for this subgroup is higher than for all participants, two of the six had simple passwords which ought to have been easier to remember. There was a fairly even spread of participants between those who used complicated patterns but were able to log in with a low number of attempts, and those with simpler patterns who required a lot of attempts. A more in depth look at the types of passwords used for DrawASecret is given in Section 6.3.4.2.

6.3.2 Number of failures

During the course of the experiment, many of the participants who were unable to log in were able to remember the gist of their password but not the whole thing. For instance for the DrawASecret method, most people were able to remember the pattern they had drawn, but could not necessarily get it positioned correctly on the screen in order to successfully log in. There may have been some participants who managed to log in correctly by guessing where to start rather than actively remembering.

On Awase-E, those who failed could generally remember at least two of the images but failed on the final one. Those who were not able to log in at all were those who had picked the initial images almost at random, and so had no way of easily remembering which images they had chosen.

Figure 35: Example of an Awase-E set of pass images where the user was unable to login

Figure 35 shows an example of this through the use of one of the participant's attempts at logging in. The images outlined in red in each of the attempts are images which the user missed when going through the selection screens and instead chose the 'No Images' option. It is clear from this that the user was easily about to remember the sunflower and crescent moon images, but struggled to remember the Hollywood sign or the skyscrapers (not once choosing the Hollywood sign when it was shown to them).

Several of the participants tried more than one very different drawing for the password on DrawASecret. In some cases, when they came to log in, they tried redrawing one of their original unsuccessful password attempts rather than the, often simpler, password they eventually used. Only six of the sixteen participants from Group A, and five of the 14 participants from Group B were able to use their first drawing as their password for DrawASecret, with the rest choosing simpler, easier to replicate passwords after at least one failed attempt. It is unlikely most of these participants would have ever been able to recreate their drawings due to the initial complexity, subsequently opting for simpler patterns of lines or dots. The initial drawings by participants could be very complicated

(e.g. Figure 36), meaning it was nearly impossible for the participant to redraw it within the limits of the method itself. In this case, the user was unable to confirm their initial password and so eventually chose a simpler pattern to use as their password.

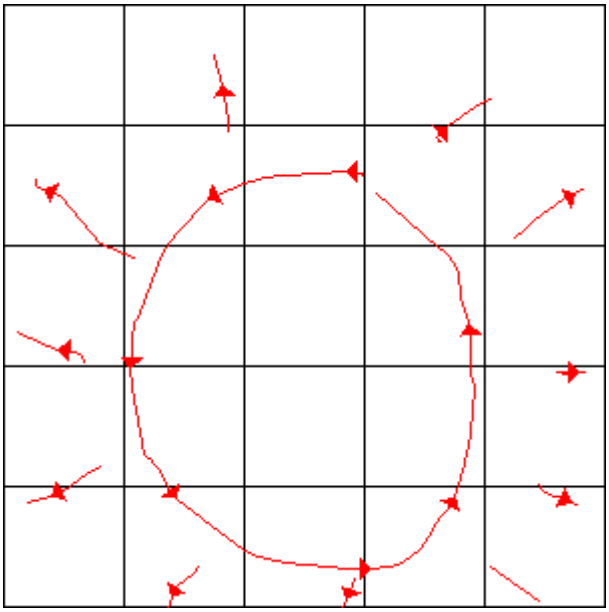


Figure 36: Example of a complex DrawASecret password

This is one of the reasons why the average number of errors for the participants using DrawASecret who were able to log in is particularly low compared to the other methods and compared to the difficulty reported by people in using the application in both groups. All of the participants (two from Group A and four from Group B) who were unable to log in using DrawASecret were able to remember the rough outline of the pattern they had actually used, but were unable to place it in the correct location on the grid, or draw it in the same scale as they had done originally. An example of this is shown in Figure 37 where the user has used a square pattern and their subsequent login attempts are all squares, but drawn in the wrong location on the grid. The last two attempts would have been correct but for the line for one of the sides of the square starting in a grid square below the one which was most likely intended, most likely due to the fat finger issue and the lack of a stylus.

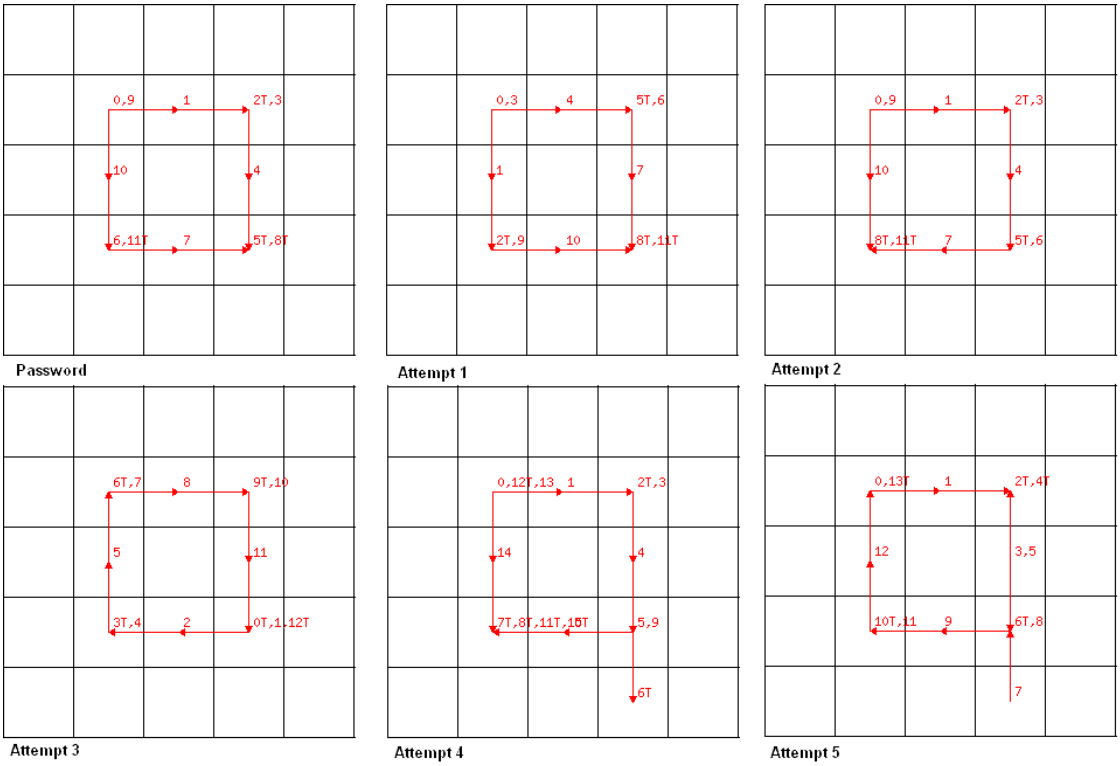


Figure 37: Example of a DrawASecret password which the participant was unable to remember and log in using

Similarly to DrawASecret, some of the people that failed to log in on PassPoints were those who had been unable to recreate their initial set of points when creating their account. This led them to choose a simple pattern of points on the screen unrelated to the background image and based on the width of the squares shown. This meant that when they came to log in they had no reference for how far apart their points had to be in order to successfully log in and so failed. One person from Group A and two people from Group B were shown to have failed to log in due to using this method to choose areas on the image. An example of this is shown in Figure 38 where the user chose five points along the bottom of the screen but without the reference grid, consistently spaced their attempts at the points too far apart to properly log in.

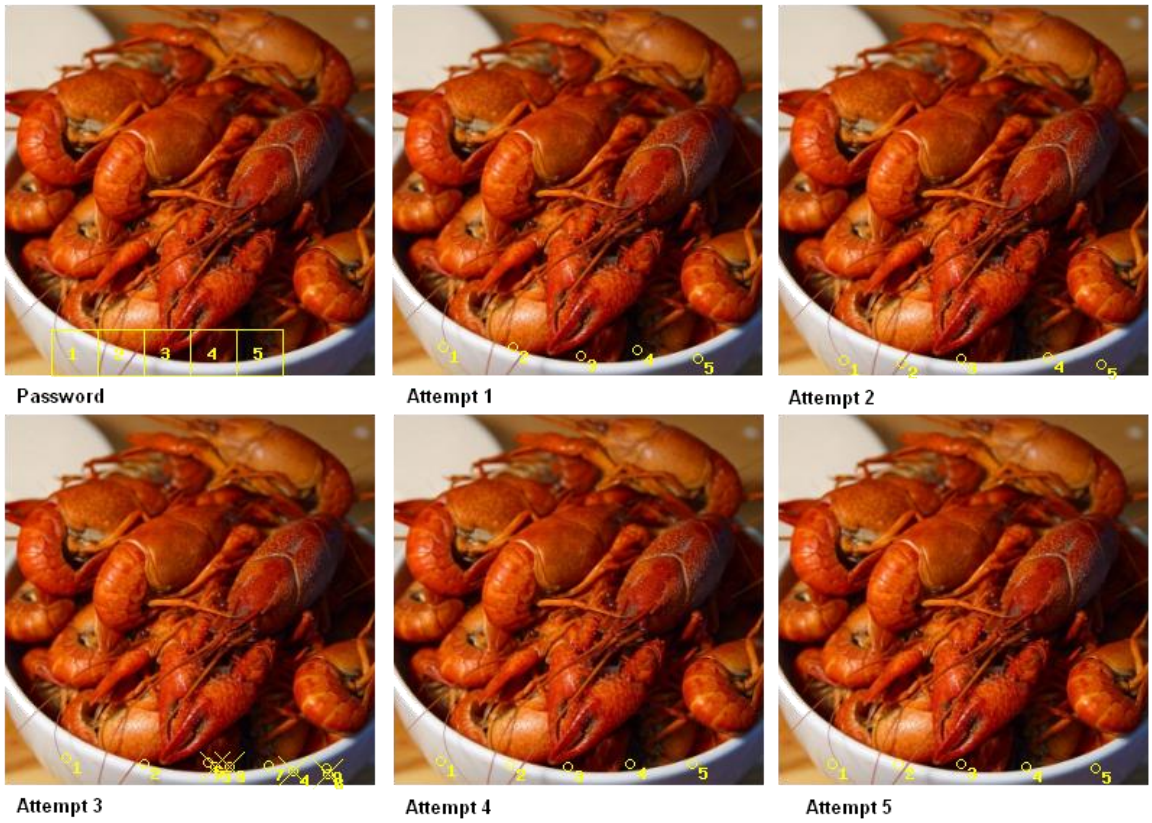


Figure 38: Example of a PassPoints password where the user was unable to remember it accurately

Two people from Group A and four people from Group B were unable to log in due to not being able to accurately place their login points on the screen despite remembering roughly which areas of the screen they had chosen. The edges of the squares were strict in whether or not a login would be allowed, and so even one pixel out would count as a failed login; some of the login attempts by participants failed by only very small margins like this.

6.3.3 Time Taken

The time it took participants to create the password on each of the applications varied significantly across the three methods. Awase-E took the shortest amount of time, but this did not require a confirmation of the images chosen. Had this been included it would have significantly increased the amount of time it took participants to create their password as

they may have had to search through the entire set of images to find the ones they originally chose.

Participants using PassPoints took a much longer amount of time to create their account, with the Group B average being almost three minutes. This will be due to the complexity of confirming the area choices for this method which often took users a lot of time trying to remember where the original boxes were positioned. The average number of errors is around one per participant indicating more time was spent actually using the application rather than quickly trying and failing different password attempts.

For all three methods there is a large difference between the creating account times taken when comparing Groups A and B. Group B on average took almost 50% longer than Group A for each of the applications, which helps to validate the original theory that Group A participants would be better are using the applications than Group B participants.

The times taken to log in using the Awase-E method are also relatively large, particularly when compared to DrawASecret. Overall, this would normally be expected to cause a lower usability score but this seemed not to be the case. The average time is similar to that of PassPoints which scored significantly worse in terms of usability. The difference between the two is likely to be due to the fewer errors associated with the same amount of time on the two projects. Although it took users on Awase-E about as long to log in as PassPoints, Awase-E itself takes longer to work through as a process over PassPoints. The process of choosing the images adds a delay to each step of the process which pushes the time up without people being frustrated at the system.

There were also three outliers in Group B which pushed the average completion time up significantly. If the users who made more than five failed attempts are removed, as these people can safely be assumed to simply have been guessing the password by that point, then the average drops to 51 seconds for login rather than 84 overall and 64 for those who made successful logins.

PassPoints however presented users with a single screen for them to choose their points so users who were unable to get through the login process were by through the whole process, rather than simply working through a slightly longer set of screens. This frustration along with perhaps the idea that Awase-E was easier may have led to the discrepancy in the length of time taken vs. the usability scores for Awase-E.

Overall this indicates that there is a difference between the usability of the three methods, which adds evidence towards rejecting H01, even though this was not reflected in the usability scores. Particularly in terms of the number of participants who were able to login successfully, it implies that Awase-E would be the most usable of the three methods, and PassPoints the least usable.

6.3.4 Password Analysis

6.3.4.1 Awase-E

There was a wide spread of images chosen by different participants when creating their passwords on the Awase-E method. Of the 119 images available, the most used image (that of a black cat) was used by six of the 36 participants followed by one of explosions and one of penguins each chosen by five participants. 39 of the images were not chosen by

any of the participants indicating that perhaps these images were considered unmemorable, or not interesting enough to be picked.

There was however evidence that the position of the images within the gallery when the participant was choosing their images played a significant role in which images were chosen. The gallery application used on the phone to display the images was a simple 1 dimensional scroll where the user had to scroll to the right to view all of the images.

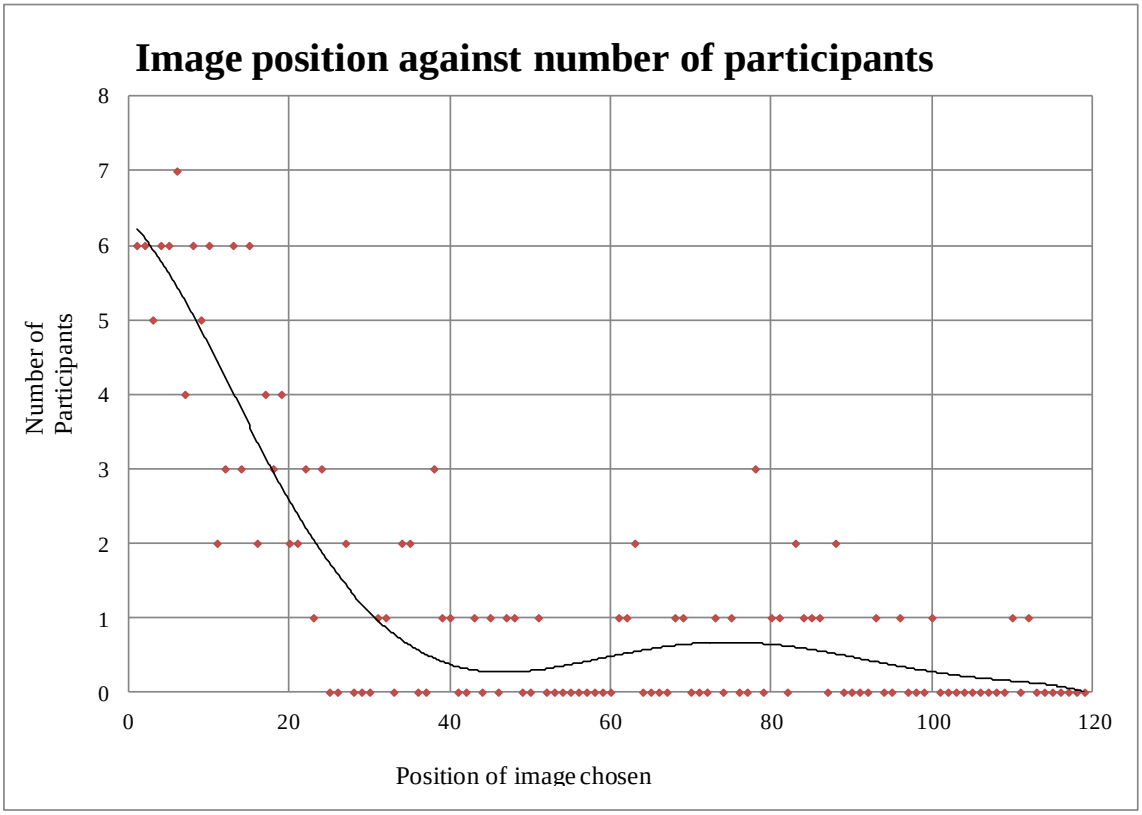


Figure 39: Graph of image position against number of participants for the Awase-E application

Figure 39 shows a graph of the position of the images chosen against the number of participants choosing an image from that position in the gallery. There is a clear trend for users to pick the images from the beginning of the gallery over the ones at the end, probably because these are the most obvious ones being presented to them. The

application was designed to prevent this from being a security issue by randomising the order in which images appeared in the gallery. This meant that even though people chose images near the beginning of the gallery, they would be choosing different images from each other each time. Therefore, there would be no way of knowing the order in which the images had been displayed to the user in order to make a guess at which images they might have chosen.

It does however mean that a lot of users are not seeing the full range of images available to them as pass images. Although not an issue in itself, this means they may not have chosen the most memorable images available from the set. If the user came back to the login after an extended period of time, and chose the images they feel they were most likely to have picked the first time, they may choose the wrong images due to seeing better or more relevant images from those presented on the login screen.

It is difficult to imagine how this could be resolved or improved on the device given the constraints of the screen size and number of available images to be displayed. It would be physically impossible to fit all of the available images onto a single screen so there needs to be scrolling of some form. One alternative might be to have a grid of images rather than a gallery structure for choosing the pass images. This would enable more images to be presented to the user at any given time meaning a smaller amount of scrolling would be required to see the entire set. This might encourage users to look through the entire set of images. It would also give the users a view of the images which was more relevant to the login process, rather than showing them the larger gallery sized images when creating the password and only a thumbnail sized image when asking them to recognise it to log in.

6.3.4.2 DrawASecret

When creating their passwords using DrawASecret, users were required to have a password length to incorporate at least 3 distinct squares. This could be in the form of a line through three difference squares or three distinct points, or a combination of the two. Many of the passwords chosen by users were close to this limit and even those that were not showed distinct signs of symmetry. Symmetry within DrawASecret means the password is not as secure as it could be and is not using the full benefit of the added security gained through the use of a graphical password. An attacker would only need to guess one half of a presumed to be obvious password in order to authenticate as the real user.

Additionally, even amongst the relatively small sample, many of the passwords chosen were similar, if not the same, between entirely separate participants. Two participants drew squares around the outside of the grid, both starting in the same square (top left) and two more drew a square around the centre square. Additionally, 20 of the passwords consisted of only a single stroke (defined as the movement of the finger across the screen without lifting), of varying length and 10 more of the passwords were a series of points on the screen. Four of the passwords were simple lines either horizontally or vertically and two participants chosen identical passwords in an L shape from the top left square to the bottom right. This supports the work done into patterns within DrawASecret passwords done by Nali & Thorpe (2004) where they found that 80% of participants created passwords using only one stroke.

		Group A	Group B
Number of exactly symmetrical passwords	Vertically	6	6
	Horizontally	2	4
	Diagonally	6	5
	Total distinct	9	12
Stroke averages	Average number of strokes	2.06	3.16
	Average number of squares used	9.83	7
	Average Stroke Length	8.12	4.29
Average length of database password		13.94	13.33

Table 22: The number of symmetrical drawings and average stroke details for DrawASecret passwords for each group

Table 22 shows some of the analysis of the drawings used for the DrawASecret passwords. In this instance, for a drawing to be symmetrical it had to be completely symmetric, rather than with an element of symmetry (for instance, a square with an off centre dot would not count here, although the square would be a symmetrical element in the drawing). Some of the passwords drawn were symmetrical in several different ways, such as a square being symmetrical in all three ways of horizontal, vertical and diagonally and so were counted multiple times. Participants in each of the groups were quite similar in terms of the symmetry with at least half of the participants from each group choosing a symmetrical password.

The average number of strokes does show some differences between the two groups. The participants in Group A used on average one fewer stroke than those from Group B, but made the stroke longer. Through this, the two sets of participants ended up with similar average passwords lengths, although the ones from Group A would be more secure.

Several of the participants from Group B used only a set of points on the screen as their password which would have pushed the averages towards this effect.

19 of the 36 participants (nine from Group A and ten from Group B) were symmetrical at least at some point around a point or line indicating their passwords would not be secure enough. This further supports Nali & Thorpe (2004) by showing that that many users choose very symmetrical passwords ones centred in the grid itself. Furthermore they found that 80% of their participants used short passwords consisting of between one and three movements. In this experiment, 15 of the 18 Group A participants used passwords of this length and 10 of the 18 participants from Group B, with eight and 12 participants respectively from groups A and B choosing passwords with only a single stroke.

6.3.4.3 PassPoints

As shown in Table 23, 13 of the 16 available images were used at least once by one of the participants during the experiment. The most popular image was the one of the Golden Gate Bridge which was used by eight participants, followed by the Hieroglyphics used by five. This may indicate an assumption by participants that they would be able to choose their PassPoints easily on an image where there are obvious features (such as the bridge towers or individual glyphs).

Image Description	Students	Public	Total
Antique Car	0	1	1
Beer Bottles	0	0	0
Burger and Chips	1	1	2
Champagne	1	0	1

Crawfish/Lobster	1	1	2
Golden Gate Bridge	5	3	8
Hieroglyphics	3	2	5
Hot Air Balloons	2	2	4
Office Space	2	0	2
Prickly Pears	0	0	0
Purple Flowers	0	1	1
Sailing Boat	0	3	3
Ships Wheel	0	0	0
Shopping Centre	1	0	1
Sunflower	1	3	4
Walt Disney Concert Hall	1	1	2

Table 23: Background images used for PassPoints passwords

Five of the images accounted for 2/3 of the choices made (24/36). In order of most chosen these were the Golden Gate Bridge, Hieroglyphics, Hot Air Balloons, The Sunflower and the Sailing Boat. These images are all ones with specific objects in the photos with reasonably distinct areas to choose (for instance the tops of the towers on the Golden Gate Bridge). The busier images such as the shopping centre, or office desk were chosen rarely if at all.

Some of the users did not use the background image as a basis for the points they chose as their password, such as shown in Figure 40.

Figure 40: Example of a set of Pass Points not related to the background image

When asked afterwards about this, they replied they had not realised or considered using the image as a cue for their points so this could be solved through improved instructions or by using a test run through the application to teach them how to use it as mentioned previously. Those users that did not use the background image chose easy to remember squares like all of the corners and the centre, which again would be easy for an attacker to make a reasonable guess.

This shows that even people not familiar with the application itself, at least initially, are instinctively able to choose images which would reduce their mental load when coming to log in or simply remember them. Although good from a usability standpoint, this does significant reduce the security of the method as a whole as it would be fairly easy for an attacker to guess the most likely places to pick to try to gain entry. Although no two participants picked the exact same points, certain areas of the images chosen were repeated multiple times.

Within the images chosen, there was evidence of 'hotspots' appearing where multiple participants chose the same points on the same image. As with DrawASecret, this is a security issue in that an attacker could make a reasonable guess at the likely points on the screen and be able to guess several of the points in this manner.

With PassPoints however, the user does have to select the correct background image. If an attacker were to choose the wrong image then they would never be able to log n, regardless of whether or not they might have picked the correct points on the screen. Similarly if they were to choose the wrong image then it is in turn highly unlikely that an attacker would be able to guess the correct points since the points would no longer match up to the background image.

6.4 Analysis of comments and qualitative data – Dataset 3

This section looks into the qualitative data collected during the course of the experiment. This was in the form both of comments made by the participants verbally during the course of the experiment, or written in the additional comments section on the questionnaire.

6.4.1 Awase-E

Out of Group B, only two people failed to login using their created password and only one person from Group A. For the most part, this was the preferred application for participants from both of the groups. Only three participants from each of the groups specifically preferred one of the other applications. The people who struggled with the application were mostly people who chose the images nearly at random, rather than those who specifically chose images either liked or were able to associate with (for instance, one

person chose two images related to San Francisco as he had recently been there on holiday).

A couple of people commented that it would be easier to remember the images if they had been able to choose their own images as well as the pass images provided. Although not feasible in the usability study, this could be possible using an application on the participant's own phone if they store any personal images on it. However in order for this to then be secure, a substantial number of other personal images would have to be added in order to create a sufficiently large distraction set of images for the login process.

6.4.2 DrawASecret

Two participants were unable to log in from Group A and four from Group B and on average this came out second most popular of the three methods participants chose. Two people from Group B and two Group A chose this as their favourite, or joint favourite of the applications. This type of application favoured people who have an inclination for remembering patterns, and those that excelled at using this application tended to do worse on Awase-E remembering set images. Having said that, a substantial number of people were able to remember the pattern they had drawn, but were unable to replicate it accurately enough to be able to log in. For example, several participants chose to try to draw their initials. When recreating, although they drew the same pattern they would often draw it at a different size or starting in a different grid and so would be unable to replicate it in the create account process, let alone when logging in. This sort of issue led many people to choose very simple patterns (a straight line, or small set of points) in order to get through the process. This somewhat defeats the purpose of the application and is a good example of why usability is an important factor in any security system.

This also explains the large difference between the times taken for participants to create their account on DrawASecret compared to the short amount of time taken to log in. If a user initially tried to use a very complicated password and was unable to recreate it at the creation stage of the process they would normally choose a far simpler password which they were able to recreate. This meant that when it came to logging in, the password they had to remember was quite short and very easy to remember, although insecure. This is shown in Table 21 where the average time taken for a successful login was only 7.6 seconds for Group A and 9.7 for Group B. This short time would by itself indicate that the login system was very easy to use.

However, when taking into account the frustration and time taken over the create account phase (an average of 77 seconds for Group A and 135 for Group B to create accounts) the lower scores for DrawASecret as a whole can be explained. This is also the reason why there are proportionally so few errors for the successful DrawASecret (0.2 for Group A and 0.4 for Group B) since the people that were able to remember their passwords were often, but not always, the ones who had chosen very simple and quick passwords which would have been quite difficult to get wrong.

Those in Group A may have had a slight advantage over the often older people in Group B due to an increased familiarity with this sort of login method. For example, Google android phones have for a while had a pattern based unlock screen as a default feature, although it needs to be actively chosen by the user. Despite this possibly higher familiarity though, Group A participants were no more inclined towards choosing DrawASecret over the others. This could be because although they may be more familiar with the Google

method, the differences between DrawASecret and the Android PatternLock method were distinct enough to balance this out by causing more failures.

Future versions of DrawASecret would probably have to involve the matching of patterns themselves (so an S shape) rather than a strict adherence to a grid structure, which was found to be inadequate for drawing curved lines. There were also issues with the grid structure itself. Participants often strayed too close to the grid lines, to the point where they accidently crossed over into a neighbouring grid, without realising. This caused the password created to involve a line into the neighbour grid and so made it difficult to replicate. This may also be solved with more explicit instructions and training.

6.4.3 PassPoints

Seven people from Group B were unable to log in using their password on this application, and three from the Group A. It was the method most disliked by the majority of the participants, often scoring 1-2 points below Awase-E. Conversely, those that did understand the concept preferred it to the other methods. A substantial number of people didn't realise that the background image could be used to place the grids on the screen and so opted for simple to create passwords such as using the first five squares across the top of the screen. These however are very difficult to recreate when logging in as the size of the squares chosen for the password is no longer visible. This left those who had used this method for the most part unable to log in, even though they remembered which squares they had chosen and so left them feeling very frustrated.

Similarly, participants were relying on the squares themselves to give a clue on the underlying grid and so line up subsequent pass areas that way. This compounded the issue of the squares "changing" to circles as people could no longer use the visual clue of how large the square was and see how far away the next point ought to be when logging in. One way to get around this in future might be instead of using squares to show the entire area available, use the circles used for logging in centred on the point where they pressed it. The square area could still be used as the margin for error, but not displayed explicitly to the user. This could create issues when repeating the account though if and when the user isn't close enough to the original point. This may put them into a different virtual grid thus stopping the repeated password from being confirmed as the same.

Another issue noted by most of the participants was the disparity between the squares used on the create account stage and the circles used when logging in. Although this was part of the design (the smaller circles being able to go anywhere inside the chosen pass areas) it confused a lot of people.

This may be partly to do with the instructions not being clear enough on ideas and concepts behind the method. This initial confusion may have put some participants off when they first started logging in and may have contributed to being unable to log in. Some participants were able to remember where the squares would be placed on the screen, particularly with reference to the other squares, but were unable to map this to putting the marker circles in the correct place for logging in.

Participants also had issues with the underlying grid structure. Many were unable to put the square onto the image exactly where they wanted it. This was in part due to the grid

being set, so the possible squares rarely matched perfectly with the image in question, and in part due to the innate inaccuracy of using a comparatively large finger to choose a small location on the screen. Whilst many participants either requested or pointed out the need for a stylus type device, one was not available which would work on the capacitive touchscreen used for the experiment. Without a precision implement, PassPoints becomes substantially more difficult to use on the smaller screen, and this is reflected in the lower scores relating to this point on the questionnaire. The issue of the stylus was most obvious on PassPoints, but would have helped the drawing of images on DrawASecret as well since the participants would have had a better idea of which areas of the screen they were touching at any given time, rather than another part of their finger leaning into one of the other grids and changing their password.

6.4.4 General Comments

Many of the comments included as additional remarks on the questionnaire, or made by participants during the course of the experiment, highlighted the usability issues of all three methods. PassPoints was highlighted as being the least usable, and consequently the least liked of the three authentication methods. 15 of the 18 participants from Group B said they preferred Awase-E to either of the other two as did 14 of the 18 from Group A (with two more unable to decide between Awase-E and one of the other methods).

Six participants from Group B made specific comments about disliking PassPoints or some aspect of it, and ten of those from Group A also made comments despite in some cases giving it higher usability scores. Only two participants made additional comments favouring PassPoints, and these both related to the additional security it might offer, in part

because of the difficulty in using it. Six participants from Group B and four from Group A made negative comments about DrawASecret, with only two people making positive comments.

Only six participants made negative comments about Awase-E, three from each group. Four of these related to the images being either too difficult to remember, and in two cases that using personal images would be easier. The final two comments related to the lower level of security offered by Awase-E compared to the other authentication methods. Six participants overall specifically mentioned they enjoyed using Awase-E.

These comments indicate a substantial preference for Awase-E over the other two methods which was not shown as strongly in the usability results from the questionnaire, which contributes to rejecting the null hypothesis H01.

6.5 Other Factors

There are various other factors beyond the approximate age of participants and the order of methods that may have affected the results.

6.5.1 Gender

The design of the experiment has not taken into account any differences between the spatial awareness of male and female participants and so there are unequal numbers of these in both groups. In Group A there were only four female participants (reflecting perhaps the relatively low proportion of female students taking computer science) and so

little statistical analysis can reasonably be done to show any significant difference between the two genders.

The situation is reversed for Group B with 12 out of the 18 participants being female, and so it was possible to perform a statistical comparison between the male and female participants.

Awase-E

Criteria	C2	C3	C4	C5	C6	C11	Gen1	Gen2
Average	-0.75	-0.29	0	0.13	-0.5	0.92	0.39	-0.46
T Statistic	0.16	0.38	1	0.82	0.12	0.28	0.62	0.29

Table 24: Comparison between female and male scores for each criterion for Group B on the Awase-E method

DrawASecret

Criteria	C2	C3	C4	C5	C6	C11	Gen1	Gen2
Average	-0.5	-0.75	0.33	-0.83	-0.5	0.83	-0.39	-0.33
T Statistic	0.68	0.22	0.3	0.29	0.49	0.25	0.65	0.9

Table 25: Comparison between female and male scores for each criterion for Group B on the DrawASecret method

PassPoints

Criteria	C2	C3	C4	C5	C6	C11	Gen1	Gen2
Average	-0.33	-0.75	-0.42	-1.21	-1.17	1.5	-0.33	-1.17
T Statistic	0.73	0.13	0.49	0.11	0.09	0.16	0.67	0.13

Table 26: Comparison between female and male scores for each criterion for Group B on the PassPoints method

Tables 24 to 26 show the differences in means between the male and female groups from Group B for each of the criteria. Although the average difference seems to be negative for most of the criteria (indicating that women tended to score lower than men), a standard t-

test shows no significant difference between the two sets of results since none of the results are below 0.05. This implies there is no difference in the results between men and women on the tests performed.

Women consistently scored more than men on the C11 criterion indicating that perhaps they felt the applications were less suitable to a mobile device than men did. This was the only criterion on which women scored higher than men for each of the methods but still did not indicate a statistical difference. There was also no particular difference between men and women on the DrawASecret method.

6.5.2 Past Experience

Another factor that may have affected the results is past experience of mobile phones, specifically touchscreen ones. Again this was not taken into consideration in the design of the experiment, but each participant's experience with smartphones was recorded as part of the questionnaire. If users have used or own a touchscreen phone they grasp the concept of using the methods in the experiment significantly faster than if they have never used one before. This is likely to apply more to participants from Group B than Group A. Whilst not all of the participants in Group A owned a touchscreen/smartphone even those that said they had never used one were considerably more familiar with the concept than those from the older population sample. Not being familiar enough with the technology would cause a significantly steeper learning curve and so would likely lead to lower scores for the different criteria. The following section contains tables showing the differences between participants with different levels of experience on smartphones.

The three possible options were if the participants had never used and never seen, had some experience of, or owned a smartphone. Several of the participants did not personally own a smartphone or device, but knew other members of their family or friends with them, and so had some experience of how they would be used. This data was taken from an open question on the questionnaire, rather than a multiple choice option.

Three comparisons were made:

Comparison (i): Between participants who own a smartphone and those who have experience of one.

Comparison (ii): Between participants who own a smartphone and those who have never used one.

Comparison (iii): Between participants who have some experience of a smartphone and those who have none.

Group A

Tables 27 to 29 show the difference in usability scores for participants with different amounts of previous experience of smartphones from Group A separated by the methods. Each of the possible comparisons between participants with no experience, some experience and a lot of experience are shown, along with the t statistic to show whether or not the difference between the two sets of data is statistically significant. A t statistic below 0.05 would indicate there is a statistically significant difference between the scores from the two sets of participants.

Comparison		C2	C3	C4	C5	C6	C11	Gen1	Gen2	Average
(i)	Average	1.3	-0.6	-0.5	-0.2	-0.6	1.7	0.27	-0.2	0.16
	T Statistic	0.15	0.01	0.01	0.82	0.17	0.62	0.86	0.57	N/A

(ii)	Average	0.47	0.62	0.33	0.35	0.57	0.03	0.88	1.13	0.55
	T Statistic	0.54	0.36	0.37	0.45	0.48	0.93	0.26	0.15	N/A
(iii)	Average	-0.8	1.17	0.83	0.5	1.17	-1.7	0.61	1.33	0.39
	T Statistic	0.35	0.11	0.04	0.5	0.13	0.63	0.7	0.11	N/A

Table 27: Comparison of average scores for criteria between participants with different experience levels of smartphones for the Awase-E method for the Group A

Comparison		C2	C3	C4	C5	C6	C11	Gen1	Gen2	Average
(i)	Average	1.5	0.1	-0.4	0.35	0.6	-0.6	-0.1	-0.3	0.15
	T Statistic	0.12	0.92	0.04	0.645	0.43	0.46	0.95	0.62	N/A
(ii)	Average	0.33	0.18	0.27	0.27	-0.1	0.23	0.1	0.67	0.25
	T Statistic	0.69	0.7	0.34	0.69	0.91	0.75	0.9	0.32	N/A
(iii)	Average	-1.2	0.08	0.67	-0.1	-0.7	0.83	0.17	0.92	0.09
	T Statistic	0.24	0.93	0.03	0.91	0.39	0.31	0.86	0.15	N/A

Table 28: Comparison of average scores for criteria between participants with different experience levels of smartphones for the DrawASecret method for the Group A

Comparison		C2	C3	C4	C5	C6	C11	Gen1	Gen2	Average
(i)	Average	0.4	-0.5	-0.4	-0.8	-0.3	-1.2	0.47	-1	-0.4
	T Statistic	0.58	0.53	0.6	0.69	0.82	0.57	0.56	0.13	N/A
(ii)	Average	0.4	-0.2	0.1	-0.1	0.03	-1.2	1.47	0.72	0.15
	T Statistic	0.5	0.77	0.88	0.87	0.97	0.17	0.17	0.45	N/A
(iii)	Average	0	0.33	0.5	0.67	0.33	0	1	1.67	0.56
	T Statistic	1	0.62	0.51	0.74	0.8	1	0.34	0.07	N/A

Table 29: Comparison of average scores for criteria between participants with different experience levels of smartphones for the PassPoints method for the Group A

For Group A, some of the comparisons between users with different levels of experience were shown to be statistically significant, as is shown by the t-test scores below 0.05 in the tables above. Specifically, these are C3 and C4 for Awase-E when looking at comparison

(i), along with C4 for Awase-E for comparison iii. Using DrawASecret, only C4 for comparison (i) was shown to be statistically significantly different and only C6 for comparison (ii). There were only two users from Group B who were classed as only having experience of a smartphone which makes the comparison between this group and the other groups not as statistically valid due to the low group size.

Other than these comparisons, only C6 for PassPoints for comparison (ii) was shown to be statistically significantly different. Therefore, as with Group B, there cannot be said to have been an effect on the usability results from the amount of experience the participants had of smartphones beforehand. There are indications that there is a slight improvement in both comparisons of users with no experience against those who own a device or have some experience of one.

It was shown in Section 6.1 that there was no significant difference between the three methods in terms criteria scores. As such, the results from the different methods can be averaged to produce scores for the differences between participants with different levels of experience of smartphones. The results of this are shown in Table 30.

	Average	T Statistic
Comparison (i)	-0.03	0.87
Comparison (ii)	0.35	0.1
Comparison (iii)	0.32	0.03

Table 30: Overall average differences between participants with different experience levels of smartphones for the Group A

The T statistic shows that only comparison (iii) was statistically significantly different. These results do fit with the expected differences between participants with different levels of experience.

Group B

Tables 31 to 33 show the difference in usability scores from the participants in Group B for the Awase-E, DrawASecret and PassPoints methods respectively. As before, a t statistic below 0.05 implies the comparison is statistically significant.

Comparison		C2	C3	C4	C5	C6	C11	Gen1	Gen2	Average
(i)	Average	-1.5	-0.3	0.14	0.46	-0.2	0.39	0.33	-0.2	0
	T Statistic	0.24	0.5	0.84	0.57	0.6	0.65	0.74	0.76	N/A
(ii)	Average	0.14	0.07	-0.3	0.57	0.43	0.14	1.29	0.57	0.37
	T Statistic	0.74	0.88	0.66	0.47	0.37	0.86	0.17	0.3	N/A
(iii)	Average	1.11	0.41	-0.4	0.11	0.61	-0.3	0.95	0.75	0.41
	T Statistic	0.18	0.24	0.43	0.91	0.23	0.77	0.42	0.29	N/A

Table 31: Comparison of average scores for criteria between participants with different experience levels on smartphones for the Awase-E method for the Group B

Comparison		C2	C3	C4	C5	C6	C11	Gen1	Gen2	Average
(i)	Average	-0.3	0.09	0.21	-0.4	0.64	0.79	-1.2	-0.3	-0.1
	T Statistic	0.86	0.93	0.56	0.64	0.43	0.33	0.25	0.84	N/A
(ii)	Average	0.14	-0.4	0.07	0.64	1.57	1.57	0.67	0.86	0.64
	T Statistic	0.92	0.65	0.82	0.61	0.3	0.13	0.54	0.5	N/A
(iii)	Average	0.39	-0.5	-0.1	1.09	0.93	0.79	1.9	1.16	0.7
	T Statistic	0.77	0.5	0.71	0.33	0.47	0.48	0.11	0.38	N/A

Table 32: Comparison of average scores for criteria between participants with different experience levels of smartphones for the DrawASecret method for the Group B

Comparison		C2	C3	C4	C5	C6	C11	Gen1	Gen2	Average
(i)	Average	-0.6	-0.3	-0.6	-1.4	-0.3	-0.4	-1	-1.9	-0.8
	T Statistic	0.58	0.65	0.53	0.17	0.76	0.73	0.42	0.08	N/A

(ii)	Average	-0.4	0	-1	-1.3	-0.1	-0.7	-0.7	-0.4	-0.6
	T Statistic	0.7	1	0.29	0.2	0.89	0.54	0.38	0.65	N/A
(iii)	Average	0.21	0.27	-0.4	0.07	0.18	-0.3	0.31	1.45	0.22
	T Statistic	0.81	0.69	0.49	0.94	0.84	0.73	0.79	0.22	N/A

Table 33: Comparison of average scores for criteria between participants with different experience levels of smartphones for the PassPoints method for the Group B

None of the differences between criteria were statistically significant, however there are indications that there are differences in the scores given to the criteria for the different levels of past experience but the order it outputs is Seen > Owns > No experience. Intuitively this does not make sense, since those who own a smartphone ought to be the most familiar with the technology, and so this result is likely to be a statistical anomaly. People with no experience were shown to score the methods lower than those with some experience, although again this was not a statistically significant difference.

Again, since there was not shown to be any statistically significant difference between the usability scores from the three methods, the results can be combined to show an overall difference between the participants with different levels of experience of smartphones, as is shown in Table 34.

Overall Averages

	Average	T Statistic
Comparison (i)	-0.3	0.09
Comparison (ii)	0.44	0.02
Comparison (iii)	0.12	0.56

Table 34: Overall average differences between participants with different experience levels of smartphones for Group B

These show that the only significant difference between participants was comparison (ii). It would have been expected that the difference between people with no experience and people who own a smartphone would be the most likely to be different, but oddly the reverse was true. This shows there was very little consistency across the experiment, given the exact opposite of this was shown to be the case in Group A.

6.6 Threats to Validity

This section describes some of the main threats to validity for the experiment.

6.6.1 Internal Validity

Internal validity relates to the validity of the results based on the methods used in the experiment itself and whether or not the two can truly be said to correlate to one another, both in terms of the way the experiment was designed and in terms of other factors which may affect the outcomes.

6.6.1.1 Design of Experiment 1

The Awase-E method did not require participants to confirm their chosen password, as was the case with the other two methods. This meant there was no re-affirming of the images chosen within the participants' minds, and may have led to some of the issues they found in remembering them when it came to log in. This would be particularly true of the participants who originally picked their images at random, or with no particular thought to how they might remember them later. However, despite this limitation, Awase-E was shown to be the most usable of the three methods, so the introduction of this confirmation step would be unlikely to change the results significantly.

Another issue with the design of the experiment is that there is only a small allowance for training in the use of the methods. The experiment was designed purposefully so that participants would all have no knowledge of the different methods before starting. However this means that the results are not likely to be valid for the instance where a participant uses a method over a prolonged period of time. Better training would almost certainly decrease the amount of time taken to log in, as well as reducing the number of errors made whilst doing so. It is also likely that in this situation, the usability scores given by the participants would change, most likely improving, as they became more aware of how to use the methods effectively.

Similarly, testing whether a password can be considered 'memorable' would normally require a long term study to determine whether people can remember the passwords after both a significant amount of time of use, and of no use. During the experiment, a period of about 20 minutes was allowed between creation of the password and logging in. The participants were required to perform some distraction task to prevent them from simply repeating the password over and over as a way of remembering it. Nevertheless, this is still a relatively short time period between the two stages of the method. It is felt that even for the participants who were able to remember their passwords, many of them would not have been able to do so after perhaps a week of no use. This too would be likely to affect the usability scores, and so a much longer study would be needed to address this issue.

6.6.1.2 Participant Psychological Effects

The participants themselves may have biased the results of the experiment due to the way in which they approached the tasks, and how they felt they ought to complete the

questionnaire. Even for methods where the satisfaction level was quite low overall, the average scores were often positive. No average for any criterion fell below 4 suggesting that participants might have been giving high scores to help the researcher get 'good' results. This is the good subject effect whereby a participant actively tries to give the correct or a positive answer to avoid being seen as having failed, or being inferior to other participants (Rosnow & Rosenthal, 1997).

The process used to gather the data and calculate variances for the scores given by the participants for the criteria will have removed some of the bias this causes, and allowed the differences in the results to be shown up. However, with people scoring all the methods very similarly over the course of the experiment the differences between them will have been small and so less likely to be statistically significant. It was also noted that due to the nature of the sample of participants in Group B, many of them considered themselves not to be computer literate. This would have negatively affected their performance on the mobile device since the participants equated knowledge and skills on a computer to their ability to use a smartphone. They therefore assumed they were unable to perform tasks on the phone without either a significant amount of instruction or physically being told what to do. It is suspected that if they were able to get past this assumption (many repeatedly said during the course of the experiment 'I'm not very good with computers') they would have performed better. This will have been partly addressed by separating out the participants with more and less technology experience between groups A and B (the student participants in Group A were less likely to be affected by this) and so it would be averaged out across the results as a slightly lower score. Conversely, it is possible that if participants found a method easier to use than they were expecting, even assuming they believed themselves to be computer illiterate, they might have given it a higher score.

The environment within which the participants undertook the study might also have influenced their performance. The students in Group A took part in the experiment within one of the computing labs in the university near to the end of a term when coursework and other deadlines were due for many of them. These participants also had to volunteer for the experiment and in some cases were chased up which may have slightly increased a subconscious level of animosity towards the researcher. Such animosity could lower the performance of these participants compared to what they might normally be expected to do in a real world situation. Most of the participants in Group B took part in the experiment in their own homes where they would feel comfortable and relaxed and so be likely to perform better. They also would have been more willing to 'help' the researcher since he was already there and had not chased participants willing to take part.

This leads on to the fact that no experimental situation in a laboratory setting (including people's homes since this is still a controlled environment) can ever replicate conditions a user would face in the real world, as discussed in Section 4.5. The laboratory setting is a controlled environment with little or no other distractions requiring immediate attention on behalf of the participant. In a real world setting it is expected that participants would perform significantly worse due to distractions and having to multi task. The Hawthorne effect (Rosnow & Rosenthal, 1997), whereby participants behave differently in an experimental situation when compared with a real world situation could have further added to this issue and influenced the results.

An example of a real world situation would be walking whilst trying to log in. Many participants, particularly with the PassPoints and DrawASecret methods needed to rest the device on a solid surface in order to accurately (re)draw their passwords or point at the

correct locations on the screen. This would not be possible in an environment where the participant may be moving around a lot.

The motivation of the participants may have also been swayed by the novelty value, whereby they are intrinsically interested in new technologies and want to try them out. This may have contributed to the fact that only 3 people refused to take part when asked from the non-student group even though, as previously mentioned, a lot of them felt unhappy with technology on the whole. This is more likely to have been influenced by the good subject effect in wanting to help the researcher.

Many of the traits associated with volunteering (e.g. social status, religion, place of birth) will not have affected a participant's ability to use the methods (Rosnow & Rosenthal, 1997). However, traits such as gender, intelligence and even whether the participants are inclined to approach a problem in an unconventional manner will affect how they remember the passwords. Gender has already been analysed in this section due to its effect on the spatial awareness of participants. The intelligence of the participants in Group A is likely to be higher than that of Group B since by definition the students are people who have successfully applied to university whereas this will not necessarily be the case for all of those from Group B. However, this was not tested as part of the experiment so cannot be confirmed one way or the other. Unconventional thinking patterns might aid a participant trying to remember some of the patterns or images but again this was not possible to be tested as part of the experiment so it is impossible to say which participants were better able to approach unfamiliar problems beyond their actual performance in the experiment itself.

6.6.1.3 Psychological Effects Due to the Researcher

Subconscious cues may also have led the participants to behave in different ways, and could have influenced the results. A researcher can give off such subconscious clues when the participant reaches the correct answer, and the participant, possibly also subconsciously, recognises this as happening and is able to work out the correct answer. When the participant was unable to remember their password this may have enabled them to see which the correct password was and log in where they might not have been able to otherwise.

Other psychological studies have shown that the opinions of the researcher (or past knowledge), can affect the outcome of trials. Experiments were carried out where two sets of researchers were asked to perform the same experiment on a dummy variable, where one group was told the experiment ought to succeed and the other told it should fail. The experimental setups were identical and the only differences between them were the preconceptions on whether or not the experiment ought to be a success or failure on the part of the researcher. The results showed that the experiments run by the researchers expecting the test to succeed did succeed, and vice versa. There ought to have been no difference between the results of the tests but the expectation of the researcher was enough to skew the results (Rosnow & Rosenthal, 1997).

In terms of this Experiment 1, it was expected that the Awase-E method would indeed be the most usable, followed by DrawASecret and PassPoints respectively. It is possible therefore that the expectation of the researcher did influence the results of the experiment in some way causing the expected result to happen. This was avoided as much as

consciously possible, and the comments made by the participants do back up the results obtained, so this effect should not have significantly affected the overall outcome.

Teacher expectancy effects may also have played a role in how participants performed (Rosnow & Rosenthal, 1997). Due to its nature, it was expected that those in Group B would perform less well than those in Group A and so the two groups were treated a bit differently on average, particularly in relation to the amount of explanation provided. Although help was available to everybody it was only offered if asked for or if the participant was having significant difficulty with a stage and this was preventing them from moving on, and where this happened, a note was made of what help was given. This expectation on the part of the researcher may have affected how the participants approached the tasks, with the people who realised they were not expected to perform well actually not doing so. As before this effect is likely to have been quite subtle if present at all and so would not have had a major affect on the outcome of the experiment.

One of the main ways of eliminating the effects of expectations is to repeat the experiment using different researchers. This was not possible due to the constraints of the project in this case, nor was it possible to monitor how the researcher interacted with the participants. Similarly, due to the nature of the participant selection it was not possible for the researcher to not know to which group each participant belonged. Future studies could be recorded as video to enable other researchers to check and confirm the findings made.

6.6.2 External Validity

External validity relates to how well the results of an experiment, are applicable to other situations, or groups of people. In this case, this is about whether or not the experiment

would produce similar results if performed using different subsets of the population, or if a different type of mobile device had been used during the experiment itself.

6.6.2.1 Input Mechanism

Although not specifically evaluated in the experiment, it is worth noting a few points about the input mechanism the participants had to use for each method. Awase-E has a large input screen with comparatively large images on display to choose from. These take up more than 5% of the screen each, and it is difficult to imagine a participant being unable to accurately choose the image they want each time, although it might be an issue for users with a disability or those with unsteady hands.

DrawASecret again has a relatively large input area, but this will always be hampered slightly by the user's finger being in the way of exactly where it is pressing on the screen. To make any particularly complicated drawings would be difficult due to the inability to make precise movements within a specific area. Having said this, none of the participants commented that the method was difficult to use from this perspective.

PassPoints was shown to be quite fiddly for participants to use and place their locations on the screen accurately. The initial square area for the pass location often did not appear exactly where participants expected, due to the underlying grid nature of the authentication method. The grid sizes were also perhaps too fine grained to be easily usable particularly with more elderly participants with less fine motor control of their fingers. When logging in, the circle representing their attempt at their pass location was obscured by the finger of the participant in a similar way to that of DrawASecret but without the added cue of the

line being drawn, it was more difficult for participants to see exactly where they had left the circle. This is referred to as the fat finger problem.

Without using a stylus, the problems associated with accuracy on both DrawASecret and PassPoints are likely to remain. Therefore from this perspective Awase-E would be the most likely method to be the most usable because the controls do not require as high a level of precision. Previous work has shown that even small buttons can be used by the majority of people (Siek et al., 2005), however when accuracy levels are down almost to that of individual pixels, it is likely that even a stylus would not allow the user to be accurate enough for the method.

6.6.2.2 Sampling Method

Of the 18 students who initially signed up, only 9 eventually took part, the other participants having to be found through word of mouth in other lab sessions around the School. The novelty value was certainly a factor for a lot of the students who took part. Since they were more familiar with the technology, several wanted to try a newer phone than they couldn't afford, and some were interested in the actual authentication methods themselves from an academic perspective.

The participants in Group A therefore fall more into the category of volunteer subjects than those from Group B due to the method of selection. The students were asked to volunteer en mass to take part (by signing a sheet and giving contact details during a lab session) whereas those from Group B were approached individually thereby only having to either agree to take part or not take part on the spot. Volunteers by their nature are a subset of the

larger group of people available, specifically the sort of person who would be willing to volunteer for an experiment, and have the time to do so. This in itself means the sample would not be truly representative even as a sample of computing students as a whole but that is a failing in all such experiments relying on volunteers. A random sample of the population as a whole would be required to remove this bias.

6.6.3 Construct Validity

Construct validity relates to how well an experiment is able to accurately evaluate the criteria or hypothesis it is meant to be evaluating. Specifically for Experiment 1, this involves evaluating how well the criteria themselves can be used to judge the usability of an application.

The criteria used for the experiment have been shown to have some weaknesses. As previously noted, no average score for any of the criteria dropped below four despite the comments made about the difficulties of using some of the methods. Although this can be partly explained through psychological aspects, the fact remains that the criteria used did not give a fully valid picture of the usability of the methods. They were able to highlight some of the differences between methods. Awase-E scored higher than PassPoints, but by themselves, the scores for PassPoints would indicate a passing score by most standards.

As such, either the criteria, or the Questionnaire would benefit from being modified to allow more of the negative comments to be reflected in the final scores given. The Gen1 and Gen2 criteria formulated for the experiment can be used as criteria in their own right, since these showed a lot of the differences between the methods. These both related to

directly asking participants whether or not they felt the methods were usable, and whether they enjoyed using them. The criteria associated with Gen1 and Gen2 from this experiment can be referred to as C12 and C13, and can be defined as follows:

C12. The application should be enjoyable for the user to use.

C13. The application should be suitable for its intended purpose and environment.

Satisfaction is one of the three main constituents of usability so if a user does not enjoy using an application or method, then it is likely they simply will not unless it is forced on them. No such measure was directly included in the original set of criteria and so this would be a good addition.

As detailed in Section 5.3, all three of the methods did meet the objective security related criteria from the original 11. There is, however, a wide difference in the level of security offered by each of the applications. Awase-E is the least secure of the three methods implemented. The Awase-E password itself can only be four pieces of information long, rather than the far longer passwords generated by DrawASecret and PassPoints and so could be relatively easily attacked by 'brute force' by a user with access to the device. This could initially be prevented by fully implementing the C9 criterion (after a set number of failed logins the device should be locked and even the correct password disallowed until it is unlocked by a party with a higher authentication level), but even with this there is a higher chance of one of the initial attacks being successful.

The chance of an attacker gaining entry to the system would be increased if the attack used social engineering techniques to get make a guess at the password. For instance, if the

attacker knew the user was fond of cats, it is quite possible that the user had chosen all the pictures of cats from within the image set. This was shown during the experiment where the participants who most successfully used the method and liked it were the ones who decided on a specific theme of images that meant something to them and so were able to remember it better. It may not be possible to eliminate social engineering from authentication systems which allow the user to choose their own password. Further work on this Awase-E should concentrate on making the system harder for an attacker to break in.

None of the methods presented would be resistant to 'shoulder surfing' of some form, where an attacker simply discretely watches the password input by the user for later use in gaining access to the device. Awase-E is more secure than the other two in this respect, since with DrawASecret and PassPoints, once the points or pattern is stolen then an attacker would be able to gain access instantly. Awase-E introduces an element of randomness to the login, firstly in that the images are not displayed in the same order each time, and secondly by the fact that not all four images are necessarily displayed each time the user logs in. An attacker who perhaps only saw three of the images used as the password would not immediately be able to get past a screen containing the fourth image.

6.6.4 Conclusion Validity

Conclusion validity relates to how well the results of the experiment can be used to back up the claims made from the outcome of the experiment. From Experiment 1, the main threat to conclusion validity is the relatively low number of participants involved in the experiment. Each group consisted of a sample of 18 participants, and with this number of

participants it is difficult to detect small or medium scale effects between the two groups at the 95% probability level (the calculated power for this level is between 0.60 and 0.65 for Experiment 1) (Dybå et al., 2006). However, there were enough participants to reasonably expect to detect larger scale difference between the two groups across the methods, and none of these were evident in the results. The participants were originally split into two groups due to the expectation that there would be a significant difference between them in terms of the results, so again it is surprising this was not found. Clearly a larger sample size would allow for more subtle differences to be found, and this would be required for a larger study into mobile authentication usability.

The sample of the population taken as participants for the experiment may also affect the conclusion validity. As discussed in Section 6.6.2.2, this can lead to bias in the results. However, it was felt that the samples taken were a reasonable representation of the two different levels of technological ability.

Another threat to conclusion validity relates to the way in which the data was collected and analysed. A Likert scale is, by its very nature, a discrete measurement and performing mathematic tests on it is not strictly valid. This introduces another level of uncertainty into the outcomes of the statistical analysis. However, given that no firm conclusions were based solely on the results of the statistical analysis of the questionnaire data, this effect is likely to be less of an issue to the final conclusions of the study.

6.7 Conclusions of Experiment 1

H01 was shown to be partly supported for some, but not all, of the criteria. Overall the general pattern of results was that Awase-E scored better than DrawASecret and both scored better than PassPoints. This is shown in Table 14 and is supported by the much of the qualitative data collected during the experiment (Section 6.4), and the results from the logging data (Section 6.3). Since Awase-E is a recognition based method whereas the other two are recall based, this outcome is in line with previous research relating to the differences between the ease of use of recognition and recall based tasks (Tulving & Watkins, 1973, Lockhart et al., 1976).

The methods were shown to be statistically significantly different for only two criteria, C5 and Gen2, and as previously noted, C5 has to be excluded from the results. This means that Gen2 is the only criterion for which H01 was rejected. Furthermore, within Gen2, the recognition based method Awase-E was shown to score statistically significantly better than both PassPoints and DrawASecret for Group B and PassPoints for Group A. Awase-E scored better than DrawASecret in Group A, but the difference was not significant.

The null hypothesis H02 could not be rejected, and so for the participants involved, there was no evidence that experience affected the participants' usability assessments, as shown in Section 6.1. This is despite the fact that Group B was composed of participants who were substantially less familiar with touchscreen devices than those in Group A. However, the differences in the times taken by the participants, as shown in Section 6.3, does reflect the expected results that participants from Group B would be more adept at using the applications than those from Group A.

As shown in Section 6.2, some of the comparisons between methods evaluated in different positions were shown to support H03. However, the results are too varied to conclude that the order had any effect on the scores given to the methods, and so the null hypothesis H03 is true and HA3 false.

Overall the preferred application of choice by the majority of users was the Awase-E application. This is supported both by the quantitative analysis of both what the participants said and the choices they made after performing the experiment along with the results from the questionnaire. These showed that Awase-E scored consistently higher than the other two applications for all of the criteria where there was a statistically significant difference between the methods. Even where the difference in methods was not statistically significant Awase-E did score marginally higher on average than the other two methods, although this could be a statistical anomaly. Awase-E is however the least secure of the three applications based on the complexity of the password stored.

The following chapter continues with the concept of extending the applications to make them either more secure or more usable. It details Experiment 2, which was performed to evaluate the impact on usability from security improvements to the Awase-E method.

7 Chapter Seven – Rationale and design of Experiment 2

As outlined in Chapter Six, the Awase-E method was found to be the most usable of the three chosen methods from the results of the first experiment (Experiment 1). This chapter deals with potential improvements and additions to the Awase-E method intended to increase security without significantly decreasing usability. A second experiment, Experiment 2, is described, with the intention of showing which of two proposed variations to the method increases security without too much negative impact on usability.

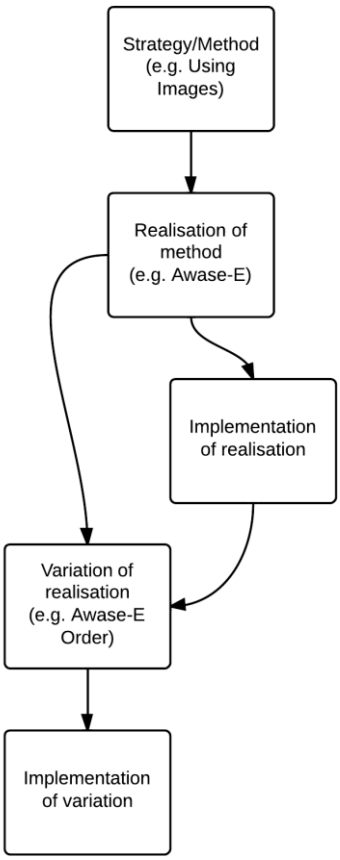


Figure 41: Breakdown of terms used to describe implementation of the original methods/strategies

The variations chosen to be evaluated in Experiment 2 are variations of the realisation of the Awase-E method used in Experiment 1. Figure 41 shows the relationship between the method realisation, implemented in Experiment 1 and one of the realisations implemented in Experiment 2, specifically the order related variation. Both of these implementations are referred to as variations for the rest of the thesis.

The two variations proposed are:

Variation 1: To increase the number of images needed for the password from four to five.

Variation 2: To require that the order in which the original four images are arranged should be an integral part of the password chosen.

7.1 Potential improved variations to the Original Methods

Following on from Experiment 1 it is apparent that none of the methods meet both the criteria for security and usability adequately. The most usable of the three methods, Awase-E, is the least secure, particularly in terms of the password space available to the user. This also means that the amount of information stored is reasonably low (only 4 names of the image files) even if they are encrypted so it would take a relatively short time for a brute force attack to discover these. Compared to the other two methods, where there are, or can be, substantially more data points for each individual login with each one encrypted in such a way as to make a brute force attack unhelpful in finding the password directly from the database, Awase-E cannot be said to be secure enough.

PassPoints however was consistently shown to be substantially the least usable of the three methods. The majority of comments made by all of the users about the method were negative either in terms of how to use it or their ability to use it. PassPoints is however one of the most secure of the three in terms of the data being stored on the device. The majority of participants in Experiment 1 used the least possible number of pass points required (i.e. 5) which is a slight improvement over Awase-E. However, the password space available is substantially more with a total of 226 possible grids which the user could have chosen. This is balanced however by the fact that since images are being used to cue the recall of users, only certain parts of the picture are likely to be chosen compared with pass points in a real situation. This could drastically reduce the number of possibilities depending on the background image used. In addition to this, each of the pass points stored in the database was doubly hashed using the method described in Section 5.3.3 as

part of the PassPoints method. This would help prevent an attacker from finding the password through a brute force attack on the data stored.

DrawASecret fell in between the other two methods in terms of usability. Although more secure than Awase-E there are still only a possible 38 possible unique data points (6x6 square grid plus a start and end point marker), although a near limitless series of lines and points is possible making the password space technically infinite. Realistically the use of a very complicated drawing would not be feasible as a password without it being too difficult for an average user to remember. From the usability perspective, it was considered to be easier to use than PassPoints but not as easy as Awase-E.

It was considered infeasible to improve on DrawASecret without pushing it to be a clone of the current Google Android graphical authentication method. This can be used on current android smartphones as an unlock screen and involves joining a series of points together. This is less secure than the DrawASecret method since it requires the pattern to be one continuous line and cannot reuse the same point, heavily restricting the number of available patterns.

The other two options are to increase the security of the Awase-E method without sacrificing usability or to increase the usability of PassPoints without sacrificing security. However with these, as with all security methods, there is a trade off between security and usability. The sections below detail some possible improvements to each of the methods and their associated benefits.

7.1.1 Awase-E

The most obvious way to improve the security of Awase-E is to increase the number of images required for a successful login to five or six. This would not represent a large decrease in the usability of the method, particularly since Awase-E is a recognition based method, but would improve the security in terms of the data points stored per user. However this would have to be balanced against the decrease in usability which is associated with having to remember more images.

Another possibility is for the user to choose more than four images but only use a maximum of four during the login procedure. This would work to help counter a shoulder surfing style attack which could be carried out, along with increasing the size of the password space. It could conceivably increase the load on the user's memory, but it is likely that having to choose additional images would mean users were more inclined to choose images with 'themes' (birds/cats/flowers etc).

One option which arises from the way participants chose images in Experiment 1 is to use the idea of themed images within the image sets available. Different people have different areas of interest or expertise, so for instance somebody who has an interest in Astronomy might be better able to tell two different images of galaxies apart than somebody who did not. They could then choose their images from a set of astronomically themed pictures from the larger set. This not only adds a layer of security to the social engineering side of passwords, but also adds an extra step to the logging in process which would need to be chosen correctly before the user, or an attacker, could log on to the system. It is however possible that this would make the system less secure overall, particularly against shoulder surfing style attacks, as an attacker could easily recognise that the user is only choosing

pictures of certain types of objects even though the user might not be picking the same pictures each time. This approach would also require a significantly larger set of images to choose from to ensure that there are a sufficient number of images from each category for the system to be secure.

Several participants taking part in Experiment 1 suggested the method ought to allow them to use their own personal photos as the pass images. Although this would increase the usability of the method by making it far easier for people to recognise images, it would also make it very easy for an attacker to distinguish the pass images from any set of stock images available to the method. Unless the user was able to provide a significant number of images from which to choose the specific one, this would not be feasible.

One of the flaws with Awase-E is that the pass images will by their nature always be displayed at some point along the login process. This means that if an attacker were to get hold of the device they could make a good guess at the correct images by cycling through the login process and watching for the images which appear the most often. Although this would be in part nullified by C9[1], even after two runs through the method it would become fairly clear which of the images were correct. If a large enough set of decoy images were used, it is unlikely the actual pass images would be obvious from cycling through login attempts. To avoid this, Awase-E can be modified so that when a user creates their account by choosing the four pass images, four additional images are chosen at random by

[1] C9: In the event of multiple wrong passwords, the application should silently block further attempts at logging in, possibly for a set period of time. If appropriate or applicable, should there be multiple repeated failures, the device should be locked until the identity of the user can be confirmed by a service provider or other authority.

the method. These images are then stored in the database along with the chosen pass images. When the user comes to log in, they are presented with the same screens as before, but in addition to one of their pass images, one of the decoy images is also shown. This means that there are eight images which would be continually repeated through the login process rather than only the correct four making it substantially more difficult for an attacker to work out which were the correct ones. In combination with C9 where they would be locked out after several failed attempts there would only be a slim chance of the attacker randomly guessing the correct images even if they were able to deduce which of the eight pass images were being repeated often enough.

Finally, the Awase-E method could be made more secure by requiring the order of the images chosen to be remembered at login along with the images themselves. This would add an extra layer of security preventing an attacker from gaining instant access even if they were able to remember the position of the images.

7.1.2 DrawASecret

DrawASecret has a variable password space, since it is up to the user to determine their own password length. However, even for relatively simple drawings, this can be shown to be larger than would be the case for a similar complexity alphanumeric password (Dunphy & Yan, 2007).

There have already been several improvements made to the original DrawASecret method proposed by (Jermyn et al., 1999) such as those discussed in Chapter Two. PassGo (Toa & Adams, 2008) is perhaps one of the most successful of these having been commercialised

in the form of GrIDsure (GrIDsure, 2011). It has also at least partially inspired commercial methods such as the graphical unlock screen on android phones and a similar application for blackberry phones called PatternLock (Tafasa, 2011). However there have been documented cases of issues with all of these types of methods related to the pattern left on a touch screen device by the user always drawing the same pattern on the screen. If the device is lost or stolen then an attack can conceivably discover the password by looking at the pattern left on the screen (Aviv et al., 2010).

Another of the main improvements to DrawASecret is the BDAS (Background DrawASecret) method (Dunphy & Yan, 2007) where a background image is added to the DrawASecret drawing screen. Several pieces of research have been conducted into this showing how people find the background image easier to cope with and easier to create more secure drawings. The original DrawASecret encouraged people to draw roughly symmetrical drawings as their passwords, and this was seen in Experiment 1 where 19 of the 36 total drawings were symmetrical about some axis (see Section 6.3.4.2). A background image did at least partially prevent this from happening as people tended to follow the image to create theirs rather than having to come up with one entirely on their own.

Due to the number of existing adaptations of DrawASecret, no additional work was proposed for DrawASecret as part of this thesis.

7.1.3 PassPoints

Security on the PassPoints method is relatively high from a theoretical standpoint. In Experiment 1, the grids used were 8 squares across, giving a total number of squares available as 226. No overlapping squares were allowed to be chosen, so the number of combinations of squares for set of five PassPoints is of the order of 10^{11} (more than $3.85x10^{11}$). However, not all of these grids are equally easy to choose. The first grid has the largest area available in which a user can click and select the grid, with the other two becoming possible only when the user chooses a point too near to one of the grid lines. Additionally, not all of the areas on a picture are likely to be used as PassPoints. Chiasson et al. (2009) showed that PassPoints and other similar click based graphical passwords are vulnerable to 'hotspots' on a background image where a lot of users chose the same or very similar PassPoints whilst ignoring the blanker or less interesting parts of the image. Whilst they suggest an improvement to help avoid this issue, it is likely that to an extent any click based scheme will be vulnerable to an attack based on this method.

This was partially confirmed by the results from Experiment 1 in that a lot of the actual points chosen by separate participants were the same points, or in very similar areas, on the screen. For example, the most popular background image used during the experiment was the Golden Gate Bridge which was chosen by eight of the participants. Of these eight, seven chose the top of the main centre tower as one of their PassPoints, and from these seven, six used the base of the same tower for another of their points. This again highlights the importance of using images with enough of these 'hotspots' to stop it from being obvious which points to choose, were an unauthorised person to attempt logging in.

Having an image too complex though would decrease the usability since it would not allow a user to remember easily which areas of the screen they chose.

A lack of usability was the main issue people had with the PassPoints method, followed by a lack of understanding of how to use it. Here, the most obvious way to improve the usability would be to increase the square size for the pass areas on the image, even though this would reduce the overall security of the password. Many of the participants in the experiment commented that the method was too fiddly to use with just fingers (no suitable stylus was available for the participants) and so the squares were too small to choose accurately. By reducing the number of grid squares across the screen to six they ought to be easier to select although this would require user testing to find an optimal size. Failing that, an option could be included allow the user to customise the size of the grids to suit their own abilities, both adding an extra element of security to the method as well as increasing the usability.

Another way to improve PassPoints would be to convert it from a recall based (albeit cued recall) to a recognition based login method. There is a substantial body of evidence that recognition is easier than recall and so changing the method would significantly improve usability (see Chapter Four).

One possibility would be to keep the original create account screens where the user chooses the squares on the screen, but when logging in the users would be presented with a series of options where they only had to choose the correct set of PassPoints from a randomly generated set. Although this would make the method significantly more usable it loses all of the extra security inherent in the original design. The login process would

become a simple 1/x chance of an attacker randomly guessing the correct password, so to have a sufficiently large password space would require an unfeasible number of options being presented.

The options presented would have to be generated randomly, and it would be relatively easy for an attacker to guess where on an image a person would be likely to select making it obvious which was the correct option. Short of manually choosing decoy PassPoints on each possible image, or using some form of pattern recognition software to discover which parts of an image were edges or points (and so most likely to be chosen) there is no way around this issue. Neither of the two options is feasible either, due to time and manpower required to maintain such a system or to the limited processing capabilities of a phone.

7.2 Selection and implementation of variations for further evaluation

The main focus of the rest of this thesis is on how to improve the security of the Awase-E method. Awase-E was shown to be the most usable of the methods and the most interesting or enjoyable according to several of the participants. Having an method users feel more inclined to want to use would mean in the longer term they would be far less likely to switch the authentication method off out of frustration. Two variations, from those discussed in Section 7.1.3 relating to the Awase-E method were chosen and then implemented and are detailed in the following sections. These variations were evaluated in Experiment 2.

Neither of the two variations have an effect on the objectively tested criteria of the original 11 (C1, C7, C8, C9 and C10), which relate to the security of the methods. The design of

the original Awase-E method addressed these criteria and these design changes are still applicable to the two new variations (see Section 5.3.1).

7.2.1 Variation 1 – Using more images

The first variation chosen was to add an extra image to the set of images to be remembered, increasing it from four to five. This is referred to as Variation 1 for the rest of this chapter. In terms of the implementation, this was a relatively minor change to the source code of the original method, and visually only required a restructuring of the way images were displayed to the user when choosing them from the gallery in terms of the layout as shown in Figure 42.

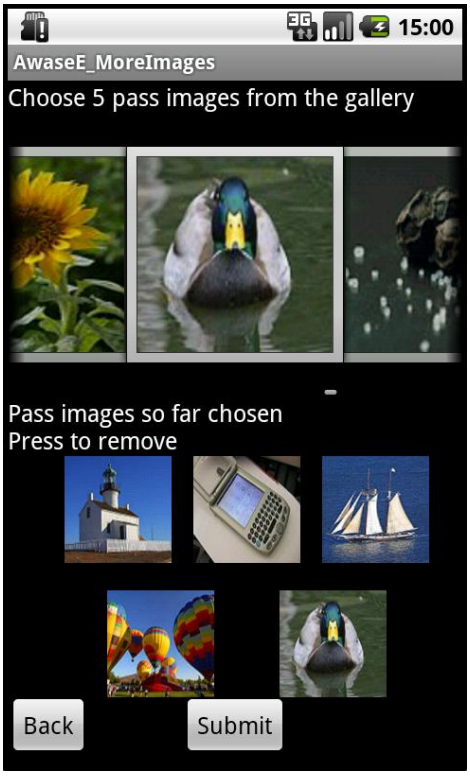


Figure 42: Screenshot of the arrangement of the five images for the Awase-E variation

Figure 43 and Figure 44 show the process diagrams for the create account and login phases of this variation, which can be contrasted with the process diagram for the original Awase-

E method used in Experiment 1 (Figure 23 and Figure 25 in Section 5.3.1). There is no difference in the processes performed by the user in logging in, or creating their account, only in the number of times the steps must be repeated to choose the correct number of images. Figure 43 shows the UML activity diagram for the create account phase of the Variation 1 method, and Figure 44 the activity diagram for the Login phase of Variation 1.

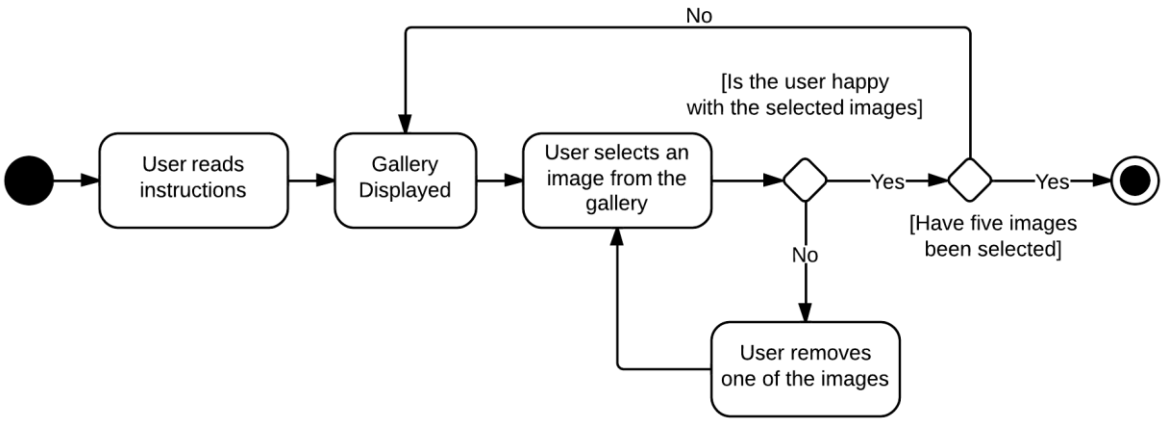


Figure 43: UML Activity Diagram for the Create Account phase of the Variation 1 method

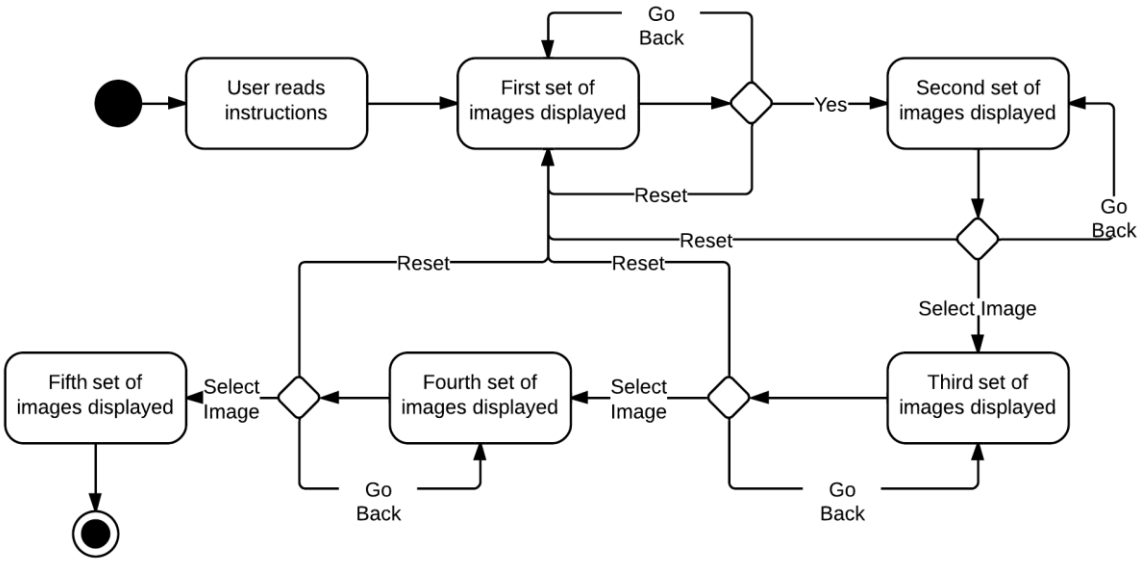


Figure 44: UML Activity Diagram for the Login phase of the Variation 1 method

For obvious reasons, this variation is more secure than the original Awase-E method since there is a $(1/13)^5$ chance of randomly guessing the correct images in order to log in rather

than $(1/13)^4$. The same set of images was used for Variation 1 as was used in Experiment 1 for the Awase-E method. Since there is no overlap between the participants involved in each of the experiments, this will not affect the results.

7.2.2 Variation 2 – order of the images

The second variation (Variation 2) to the Awase-E method was to make the order in which participants chose images part of the login process. To do this, the same process was used as before, but once the user had chosen their images, they were shown a screen allowing them to move their chosen images into a different order if they want to, or to go back and choose a new set of images. The images were arranged in a square pattern similar to that shown in Figure 45, where the images could be 'slid' onto different corners of the square.

When logging in, the user was presented with the four screens as before (as shown in Figure 24), with the images ordered randomly amongst the screens. Once the four images were selected, the 'arrange images' screen, designed to be identical to the one used for creating the password originally, was presented. This should have enabled the user to rely more on recognition memory rather than full recall making the task easier.

If the user chose the wrong images, they were not specifically informed of this, and were asked to order the images they chose anyway. If a user submitted an incorrect password they were given the option of either choosing a new set of pass images or simply rearranging the chosen images. In some cases this could mean that the user would be attempting to order an incorrect set of images, and so no combination would work without the user going back and reselecting their images.

In both the creating account and login stages, the screen for rearranging images allowed users to touch the image they wanted to use and drag it into a different position. The images 'snap to' the appropriate area of the screen depending on which of the four slots the image is closest to, rearranging the other three images in the process. This allows the user to see how the rearranged images will look without having to make an actual move each time and so they should, in theory, be able to sort the images quicker.

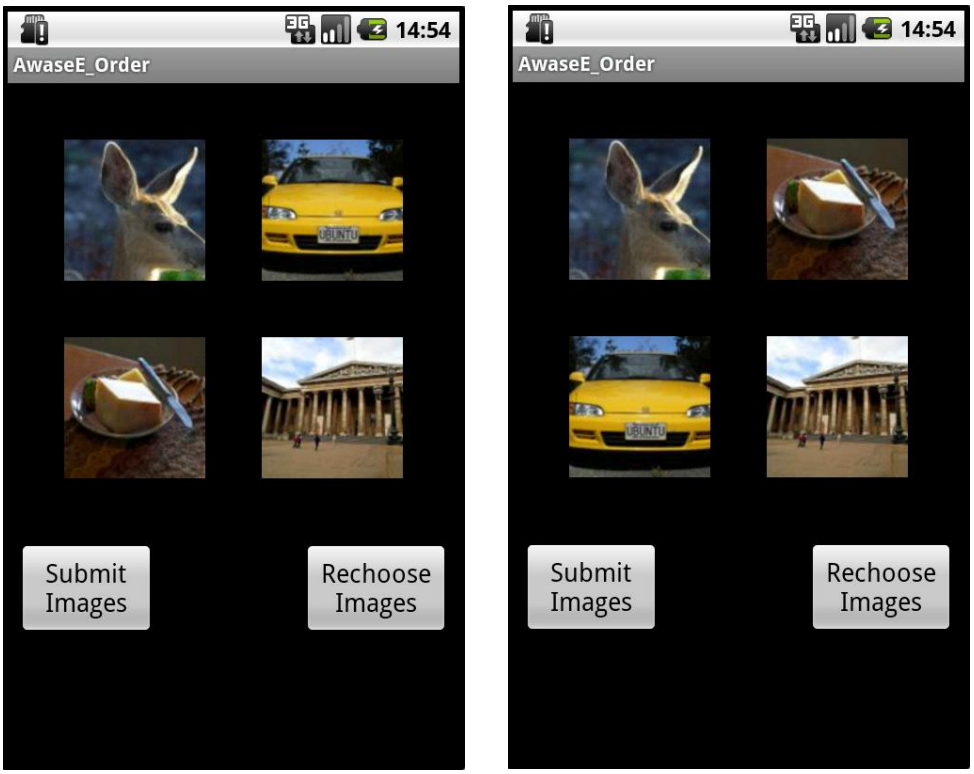


Figure 45: Screenshot of the image ordering screen for Variation 2 of Awase-E

When the background images are rearranged, they are moved linearly rather than swapping two images. So if the images are arranged with image A in position 1, image B in position 2 etc, then the process of moving image D to position 2 would mean image B moved to position 3 and image C moved to position 4 to fill in the blank left.

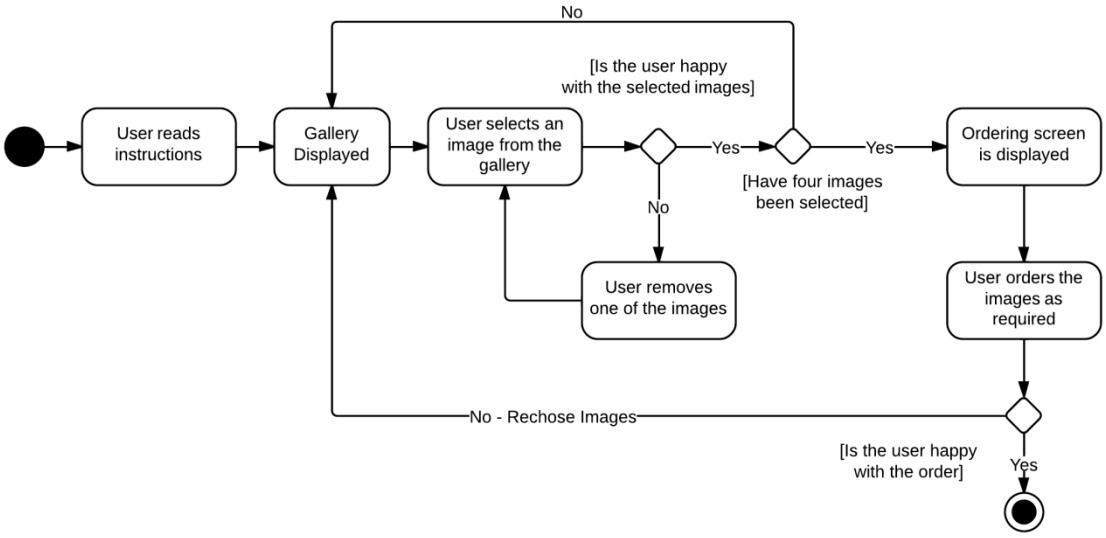


Figure 46: UML Activity Diagram for the Create Account phase of the Variation 2 method

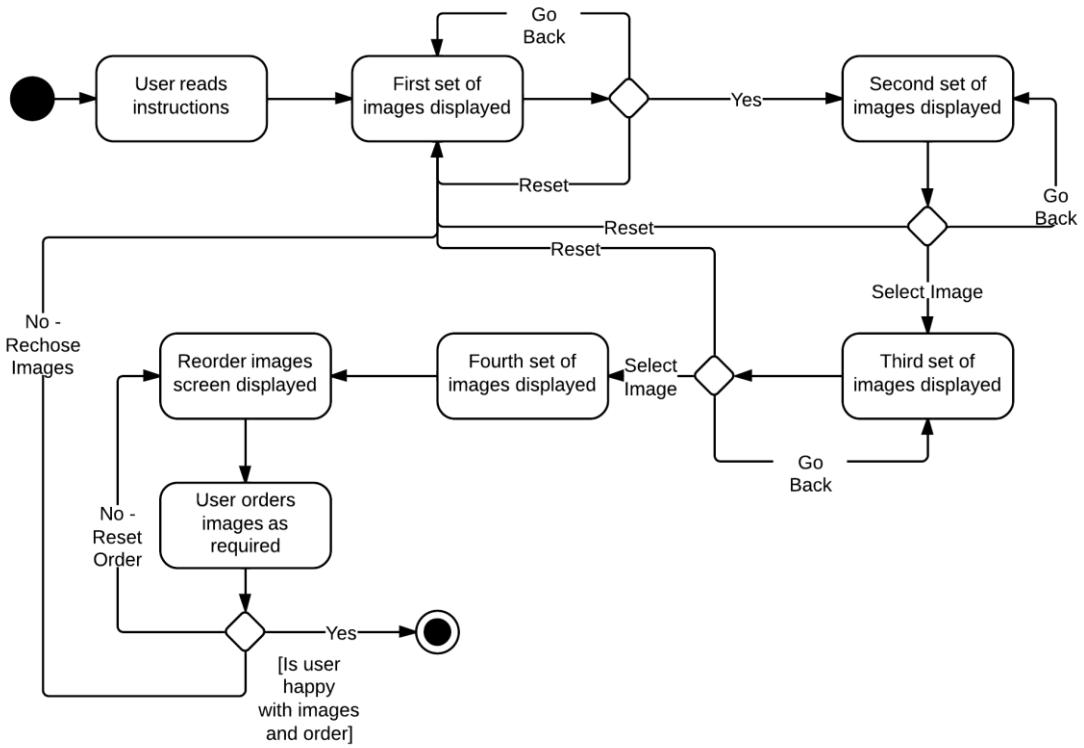


Figure 47: UML Activity Diagram for the Login phase of the Variation 2 method

Figure 46 and Figure 47 show the UML activity diagrams for the order related variation of Awase-E (Variation 2) used in Experiment 2. When compared with the diagrams for the original Awase-E method (Figure 23 and Figure 25), the additional steps required to

reorder the images are clear before the user can be classed as having successfully logged on.

In terms of the security, this variation is again more secure than the original Awase-E method since it adds an additional 24 (the number of ways of uniquely arranging the four chosen images) possible combinations to the original image choice screens. Additionally, it is a two stage authentication procedure rather than one stage which inherently adds another level of security.

To keep the two new variations independent, a completely different set of images from those used in Awase-E in Experiment 1, and reused Variation 1 in Experiment 2.

7.3 Design of Experiment 2

This section describes the design of a second experiment (Experiment 2) to compare the usability of the two variations of Awase-E and to compare their usability with that of the original Awase-E method used in Experiment 1. Many of the design elements remain the same as for with Experiment 1 to allow for a better comparison with the Awase-E results from that experiment, so only the differences are explained in detail here.

7.3.1 Hypotheses

The two main hypotheses for Experiment 2 are:

NH1) The null hypothesis is that Variation 1, where five images are required to be remembered, will be no different in terms of usability from Variation 2, where the order of

the images is required. The alternative hypothesis, NHA1 is that there will be a significant difference in usability between Variation 1 and Variation 2.

NH2) The null hypothesis is that there is no difference in usability between Variation 1 and the original Awase-E method. The alternative hypothesis (NHA2) is that there is a significant difference in usability between the original Awase-E method and Variation 1

NH3) The null hypothesis is that there is no difference in usability between Variation 2 and the original Awase-E method. The alternative hypothesis (NHA3) is that there is a significant difference in usability between the original Awase-E method and Variation 2.

Theoretically, we would expect the alternative two hypotheses (NHA2 and NHA3) to be supported since both Variations 1 and 2 require users to remember more information in order to be able to login. This increased memory load is likely to make the variations harder to use, and so users will give the usability a lower score.

7.3.2 Experimental Process

Variations 1 and 2 of the original Awase-E method were evaluated and compared with each other to determine the differences between them. Their usability was also compared with that of the original method. As with Experiment 1, Experiment 2 only dealt with the subjective criteria and so only the usability aspects of Variations 1 and 2 are investigated through the use of the questionnaire. In order to achieve a proper comparison, the only changes made to the process of Experiment 2 were the addition of either the extra image or the order requirement to the Variation. Although the changes do involve security

improvements, they do not affect the security related criteria from the original 11 criteria (see Section 5.1). As such, this experiment again was purely related to the usability of the two variations when compared to each other and the original Awase-E method.

The same questionnaire was used to collect the participants' views about the usability of variations 1 and 2 as was used in Experiment 1 (Appendix A). The same subset of the criteria, specifically those related to usability, was used as was used in Experiment 1. This was to enable a direct comparison between the variations of the Awase-E method used in Experiment 2 and the Awase-E method used in Experiment 1. The main independent and dependent variables for this experiment are much the same as those in Experiment 1, as detailed in Section 5.2.3. The only difference is in the independent variable in that only the two variations were evaluated rather than the original three methods. Similarly, the same limitations apply to the criteria as they did in Experiment 1, discussed in Sections 5.2.5 and 6.6, along with many of the same psychological factors which may have affected the results. For Experiment 2, the Gen1 and Gen2 criteria from Experiment 1 have been included with the set of original criteria and are referred to as C12 and C13.

As well as the usability scores for variations 1 and 2, other data such as the number of errors made by each participant, and the time taken to perform the tasks was also collected. As before, this was to enable a comparison of the difficulty of use of the variations and the original method. If the newer variations take significantly longer to perform the same tasks, or more errors are made by the participants then the usability of the variations must be lower. This enabled a decision to be made about which of the two variations is the most usable, in the event that the criteria alone do not show a statistically significant difference.

Experiment 2 was carried out using 18 different participants from the student population of the University. These participants were chosen in a similar manner to the original 18 student participants by asking for volunteers from students at Keele University. A different set of students from those in Experiment 1 was used to ensure that prior experience of the methods did not affect the usability scores given. This also allowed for a fairer comparison between the participants using the original Awase-E method and the two variations. Since Experiment 1 showed little or no difference between the two groups of people, students and less technologically aware people, it was assumed that a student group would be a representative sample.

The process of performing Experiment 2 was the same as that for Experiment 1, although with one fewer method to evaluate. As such, each participant was required to use both of the variations available, and fill in a separate questionnaire about each one using much the same process as illustrated in the process diagram in Figure 20. The order in which each student used each variation was alternated so that neither was consistently evaluated either first or second across the group. Average scores for each of the criteria were calculated using the same mapping of questions to criteria as show in Table 7 in Section 5.2.4. The same three Datasets were collected as described in Section 5.2.4.

The setup for Experiment 1 was replicated as closely as possible for Experiment 2. However, the cross over method from Senn (2002) used to analyse the results of the original three methods cannot be used here since the participants' evaluation of Variations 1 and 2 in Experiment 2 are different from those who evaluated the original Awase-E method in Experiment 1. Instead, a standard AB/BA comparison (also detailed in Senn (2002)) was used to compare Variation 1 and Variation 2, specifically this was a two-

sample t approach which adjusts for any period effects within the results. If this shows the two variations to be statistically significantly different, then each of the variations can be compared with the original Awase-E results using an unpaired t-test. If however Variation 1 and Variation 2 are not shown to be difference, then the results from the two can be averaged, and then compared with the original Awase-E scores, still using an unpaired t-test.

8 Chapter Eight - Experiment 2 - Results & Discussion

This section presents the results from Experiment 2. The data from Dataset 1 is first used to compare the two Variations, and then to compare these against the original Awase-E method to determine if there are any usability differences between the three methods as a whole. Dataset 2 is used to highlight the differences in the time taken to perform tasks, as well as the number of errors made by the participants using them.

8.1 Results

8.1.1 Dataset 1

Data from the questionnaires was gathered and collated in the same way for Experiment 1 to enable differences and standard errors between the three methods (the original and the two variations) to be calculated.

To show how the scores for each of the methods were distributed for each of the criteria, Tukey box and whisker plots were produced for each criterion. These are referred to as boxplots for the remainder of the chapter. Each of the boxplots shows a box representing the upper and lower 25% percentile ranges and a line representing the median scores. The whiskers represent the upper and lower ranges excluding any outliers, where an outlier is defined as any score more than 1.5 times the size of the range from the third to first quartiles (i.e. the size of the box) (Senn, 2002). These outliers are plotted separately when they occur. These plots are shown in figures 48 to 55 for the criteria labelled. Where the median value equalled one of the first or third quartile values, a bold edge is used on the box to show where the median lies. A visual inspection indicates that there are only small

differences between the scores for each of the three methods, but a full statistical analysis is presented below.

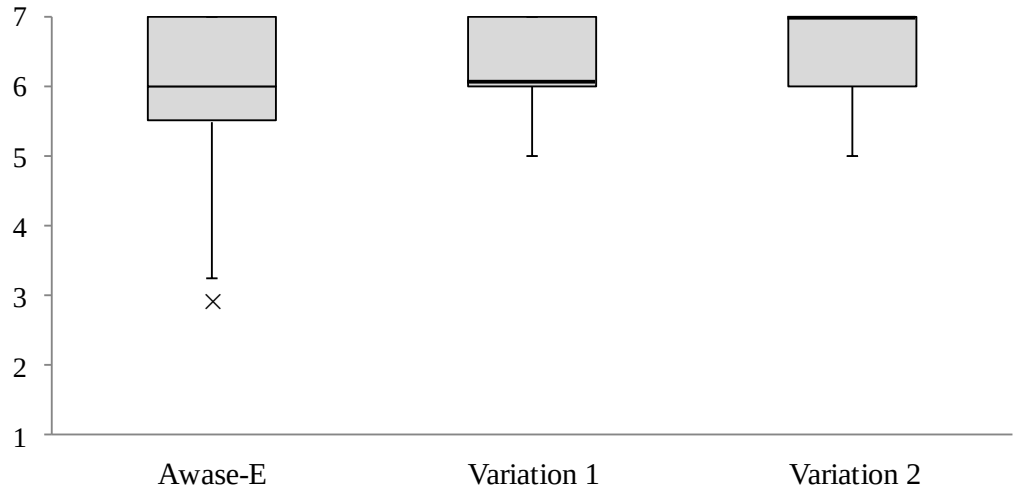


Figure 48: Boxplots showing the distribution of scores for Awase-E, Variation 1 and Variation 2 for the C2 criterion

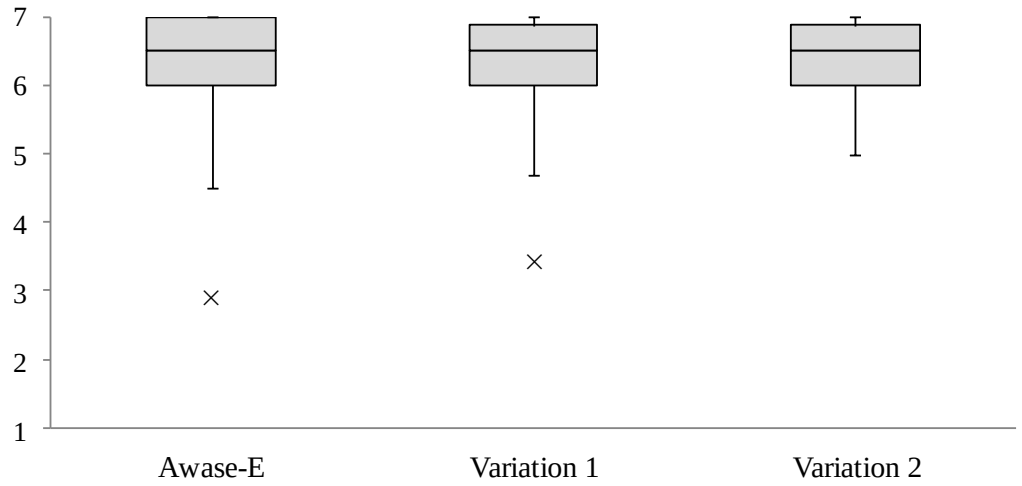


Figure 49: Boxplots showing the distribution of scores for Awase-E, Variation 1 and Variation 2 for the C3 criterion

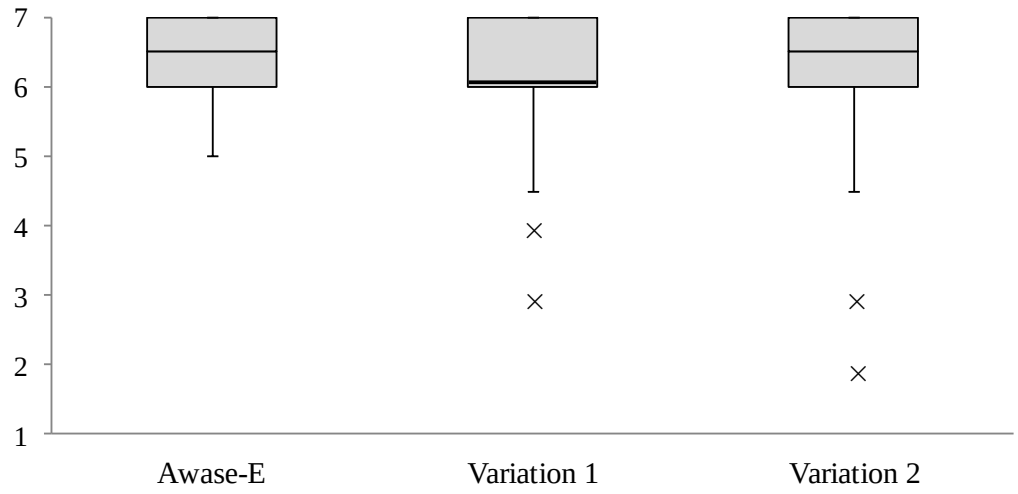


Figure 50: Boxplots showing the distribution of scores for Awase-E, Variation 1 and Variation 2 for the C4 criterion

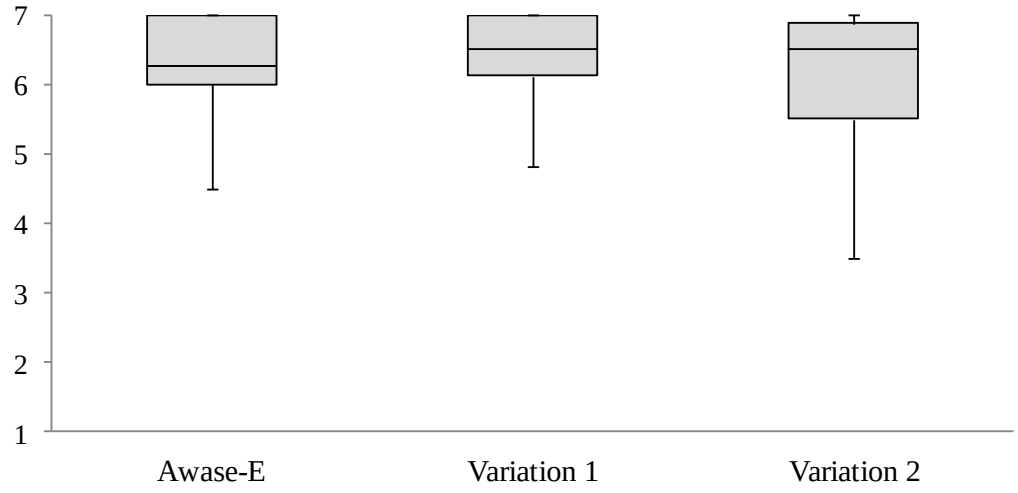


Figure 51: Boxplots showing the distribution of scores for Awase-E, Variation 1 and Variation 2 for the C5 criterion

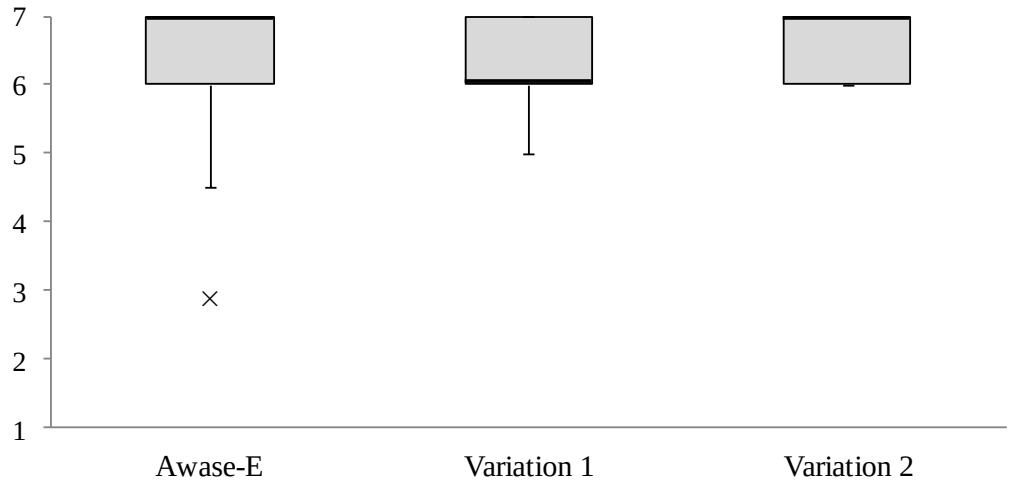


Figure 52: Boxplots showing the distribution of scores for Awase-E, Variation 1 and Variation 2 for the C6 criterion

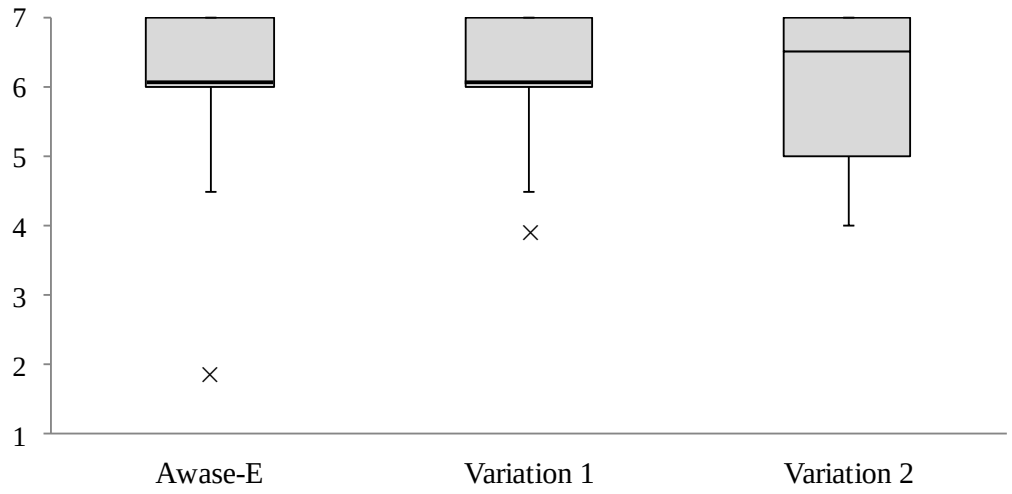


Figure 53: Boxplots showing the distribution of scores for Awase-E, Variation 1 and Variation 2 for the C11 criterion

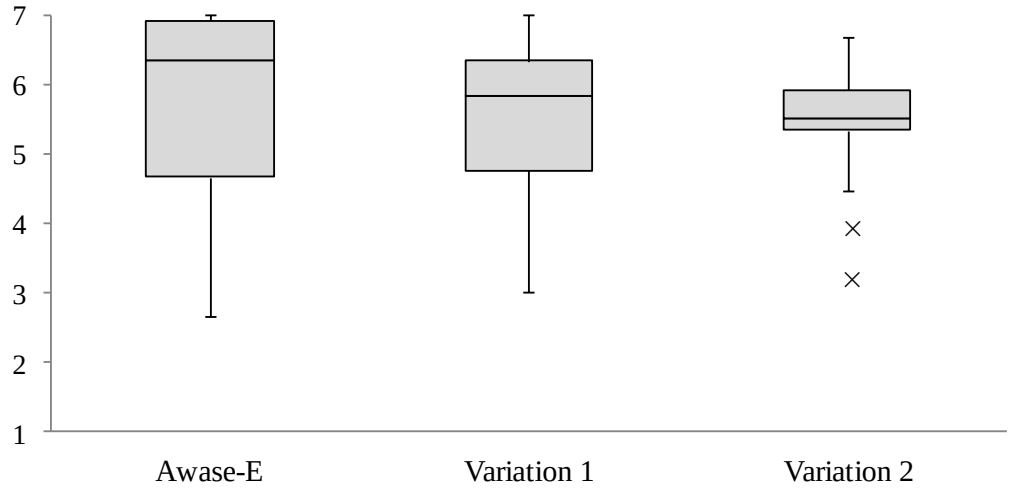


Figure 54: Boxplots showing the distribution of scores for Awase-E, Variation 1 and Variation 2 for the C12 criterion

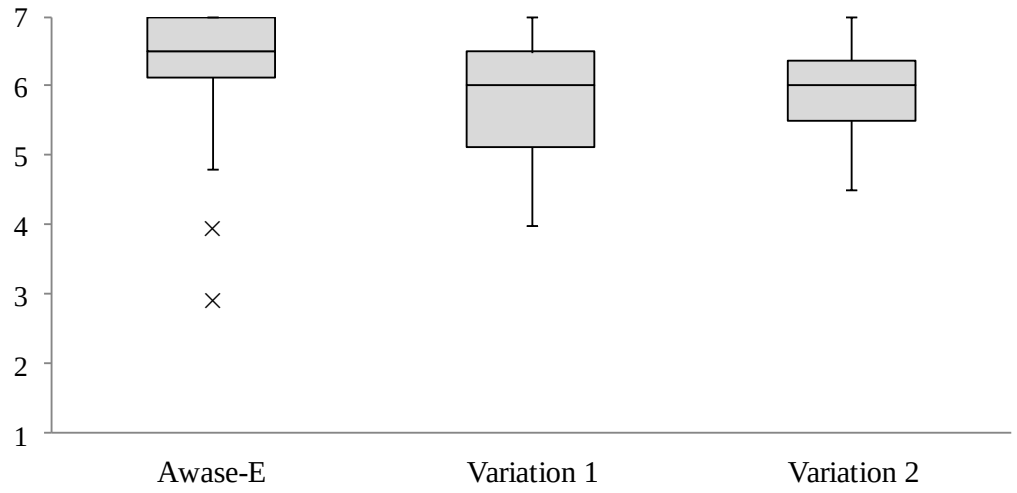


Figure 55: Boxplots showing the distribution of scores for Awase-E, Variation 1 and Variation 2 for the C13 criterion

Over the course of Experiment 2, some of the participants did not make any errors either creating an account or logging in using one or both of the new variations. However, on the questionnaire, the first question asked how they dealt with any errors they made, and so these participants answered with 'not applicable'. The average scores for the criteria C2 are therefore taken over a smaller set size than the other criteria. For Variation 1 only 11 of the 18 participants were able to give a score for this question, and 13 for Variation 2. Consequently, the number of participants for the C2 criteria is lower than for the other criteria, decreasing the degrees of freedom.

There were a significantly larger number of participants in Experiment 2 who answered 'not applicable' to this question on the questionnaire compared with that of the Experiment 1. From Group A in Experiment 1, the student group, only three of the participants did not answer, compared to seven and five for Variation 1 and 2 respectively on Experiment 2. Group B in Experiment 1 made more errors both in the logging in section and when creating the account (where the number of errors or missed attempts was not recorded) which would have contributed to more participants being able to give an answer for the first question.

To perform the crossover comparison of Variation 1 and Variation 2, whilst allowing for period effects, the results were first separated into two groups depending on which order they used the variations. The average difference between the two sets of scores for each of the criteria is calculated for each of the two groups of participants. From this, the corrected sum of squares can be calculated for each group by subtracting the mean from each of the individual differences and taking the sum of squares of the results (Senn, 2002). The

calculated method effect for variations 1 and 2 along with the confidence limits for this are shown in Table 35 along with the t and p values for the comparisons.

Criteria	Estimate of variation effect	Confidence Limits	t Value	P Value
C2	0.217	$-0.206 \leq \tau \leq 0.638$	1.161	0.276
C3	0.139	$-0.351 \leq \tau \leq 0.628$	0.601	0.556
C4	0.056	$-0.487 \leq \tau \leq 0.598$	0.217	0.831
C5	0.194	$-0.332 \leq \tau \leq 0.721$	0.783	0.445
C6	0.278	$-0.064 \leq \tau \leq 0.620$	1.721	0.105
C11	0.111	$-0.274 \leq \tau \leq 0.496$	0.612	0.594
C12	0.111	$-0.442 \leq \tau \leq 0.665$	0.426	0.676
C13	0.111	$-0.207 \leq \tau \leq 0.429$	0.741	0.470

Table 35: Method effects, confidence limits, t and p values for each criterion for the comparisons of Variations 1 and 2

None of the p values for any of the criteria are below 0.05. Moreover, the confidence limits for the method effect range between positive and negative values, so it is not possible to say whether or not the scores given for each of the criteria either improved or worsened depending on the variation being investigated. Together, this means that the difference between the usability scores between Variations 1 and 2 is not statistically significant for any of the criteria investigated. Since the two sets of participants are distinct for each of the sets of scores, an unpaired t-test was used to show whether there was a significant difference between the two sets of results.

Table 36 and Table 37 show the t-test scores for each of the criteria by comparing the scores for Awase-E used in Experiment 1 with the scores for Variations 1 and 2 respectively from Experiment 2. If the t statistic is below 0.05 then this would mean the two sets of results were statistically significantly different. For C2, an equal number of participants was taken from Experiment 1 by taking the first eleven valid scores (eleven being the number of valid results for the variations) since there was not an equal match between participants who did not give a score for C2.

Criteria	Awase-E Average	Variation 1 Average	Difference (Awase-E – Variation 1)	T Statistic
C2	5.54	6.18	-0.18	0.61
C3	6.14	6.17	0.14	0.66
C4	6.54	6.06	0.39	0.19
C5	6.32	6.39	-0.14	0.57
C6	6.18	6.28	0	1
C11	6.45	6.11	-0.11	0.75
C12	5.78	5.46	0.31	0.44
C13	6.19	5.86	0.33	0.31

Table 36: T-test scores for each of the criteria when comparing the scores for Awase-E from Experiment 1 with Variation 1 from Experiment 2

Criteria	Awase-E Average	Variation 2 Average	Difference (Awase-E – Variation 2)	T Statistic
C2	5.54	6.36	-0.36	0.33
C3	6.14	6.31	0	1

C4	6.54	6.11	0.33	0.37
C5	6.32	6.17	0.08	0.76
C6	6.18	6.61	-0.33	0.33
C11	6.45	6.06	-0.06	0.89
C12	5.78	5.43	0.35	0.33
C13	6.19	5.89	0.31	0.34

Table 37: T-test scores for each of the criteria when comparing the scores for Awase-E from Experiment 1 with the Variation 2 from Experiment 2

None of the results show a statistically significant difference between the original Awase-E and the variations in terms of the scores for each criterion.

The lack of difference between the two variations in terms of the usability scores indicates that many of the participants felt they were much the same. As such, the differences in the times taken for participants to use the variations, and the number of associated errors, is likely to be a better indicator of which of the two variations would be easier for people.

8.1.2 Dataset 2

Table 38 shows the average times for each part of the create account and login process along with the average number of errors made by the participants. As with Experiment 1, participants did not have to confirm their choice of images and so no errors could be made during the creating account process.

Method	Create Account	Login			
	Time (s)	Successful Logins	Time (s)		Number of Errors	
			All	Successful	All	Successful
Variation 1	47.9	12	110	64.6	2.5	1.25
Variation 2	66.1	14	74.6	50	2.8	1.6
Original Awase-E	57.9	17	41.6	36.2	1.2	0.8

Table 38: Average times and errors made by the participants for both of the proposed Awase-E variations as well as the results for Awase-E from Experiment 1

The times and errors for the login part is split in Table 38 between the overall average and the average only for the participants who were able to successfully log in using the respective variations. This is done to remove some of the outliers from the overall results which could skew the final analysis. Many of the users who were not able to log in took a large number of attempts before deciding to give up, which would increase the overall time and number of errors.

Table 38 also shows that it took participants longer to create their account using Variation 2 than for Variation 1. However, this did not translate into them taking longer to log in using it. In fact it took participants approximately 30% longer to complete the login using Variation 1 than Variation 2. This held true both for when all participants were considered, or when only those who successfully managed to log in using the variations were counted. This is a bit surprising as it was expected that logging in using Variation 2 would take longer due to the additional stage of having to re-order the images at the end, rather than simply adding another image from which to choose one of their pass images.

It could indicate that the addition of the extra image made the remembering process too difficult for people, whereas Variation 2 allowed some participants to use a different mental process to remember the order in addition to the images. Several participants commented that they had invented a story, or pattern, to go with the images to enable them to remember the order; one user even commented that this was a security risk since it would be possible for an attacker to make a guess at the pattern or story being used through common sense.

8.1.2.1 Number of errors

When using Variation 1, five participants were unable to log. Similarly to the original Awase-E method, most of these participants were able to remember at least two of their original images, but not the entire set. Table 39 shows the number of failed attempts for each of the users that failed along with the reasons why they failed to log in. The table shows the number of correct images chosen by the participants and the number of times they correctly chose the 'No images' option separately although these two together would count for the number of correct guesses on each attempt. The failed attempts are split between the times when the participant chose an entirely wrong image, and the number of times they chose 'No Images' when one of their pass images was in fact displayed.

Participant		Correct Images	Correctly chose no images	Images Missed	Incorrect Images
NS3	Attempt 1	1	3	0	1
	Attempt 2	3	0	0	2
	Attempt 3	2	1	1	1
	Attempt 4	4	0	0	1
	Attempt 5	3	1	0	1
	Attempt 6	4	0	0	1
	Attempt 7	3	1	1	0
	Attempt 8	4	0	1	0

	Attempt 9	4	0	1	0
NS4	Attempt 1	3	0	2	0
	Attempt 2	4	0	1	0
	Attempt 3	3	0	2	0
	Attempt 4	3	0	2	0
NS7	Attempt 1	0	1	2	2
	Attempt 2	4	0	1	0
	Attempt 3	2	1	0	2
	Attempt 4	3	0	2	0
NS12	Attempt 1	4	0	0	1
	Attempt 2	3	1	1	0
	Attempt 3	4	0	1	0
NS16	Attempt 1	3	1	0	1
	Attempt 2	4	0	0	1
	Attempt 3	3	0	0	2

Table 39: The number of attempts and the reasons a participant was unable to log in for the participants who were unable to log in using Variation 1

As can be seen from Table 39, it is clear that the majority of the participants were able to remember at least three of their chosen images on each attempt at logging in. On some of these attempts, the wrong image was chosen where a similar image (for example another image of a bird, but different to the bird chosen as part of the password) was displayed and this confused the user into choosing that image rather than selecting 'No Images'.

For Variation 2, four participants were unable to log in. Table 40 shows the number of attempts for each of these users along with the number of images chosen correctly. This table also shows the number of images that were in the correct place in the required sequence for the occasions when the participant chose the correct images, but only failed to log in due to not remembering the correct ordering for all of the images.

Participant		Correct Images	Correctly chose no images	Images Missed	Incorrect Images	Number of images in correct position
NS3	Attempt 1	2	0	0	2	N/A
	Attempt 2					N/A
	Attempt 3					N/A
	Attempt 4					N/A
	Attempt 5					N/A
NS8	Attempt 1	4	0	0	0	1
	Attempt 2	4	0	0	0	1
	Attempt 3					2
	Attempt 4					1
NS13	Attempt 1	3	0	0	1	N/A
	Attempt 2					N/A
	Attempt 3					N/A
	Attempt 4					N/A
	Attempt 5					N/A
	Attempt 6					N/A
	Attempt 7					N/A
	Attempt 8					N/A
NS17	Attempt 1	3	0	0	1	N/A
	Attempt 2					N/A
	Attempt 3	3	0	0	1	N/A
	Attempt 4	4	0	0	0	0
	Attempt 5	4	0	0	0	2
	Attempt 6					0
	Attempt 7	4	0	0	0	1

Table 40: The number of attempts and the reasons participants were unable to log in, for the participants who failed to log in using Variation 2

As can be seen in Table 40, some of the participants (specifically NS3 and NS13 here) chose their set of four images, and then stuck with those images despite repeated failures at arranging them correctly. In this situation, they would never have been able to log in. Where the number of correct positions for the images is two, the participant would have been only one action away from correctly remembering the sequence since it would require swapping the two incorrect images in order to have the correct order.

8.1.3 Dataset 3

This section looks at the qualitative data collected during the course of Experiment 2 in the form of both verbal communication and actions, as well as the comments written on the questionnaire by participants, where applicable.

At the end of the experiment, each participant was asked directly which of the two variations would prefer to use in a real environment. At this point the majority, 14 of the 18 participants, stated they would prefer Variation 2 (the order method). This confirms the results found from Dataset 2 that this was the more usable of the two variations. Strangely, four of the participants, despite saying they preferred Variation 2, actually gave Variation 1 a higher score overall on the questionnaire. This is an indication that the questionnaire may need to be further revised to fully obtain participants' opinions on the methods being investigated.

8.2 Validity

Since the process for Experiment 2 replicated as closely as possible that of Experiment 1, the threats to validity for the experiment are the same. These were discussed fully in Section 6.6. The main limiting factor in Experiment 2 is still the number of participants used. This was kept the same as for Experiment 1 for the purpose of comparing the variations with the original method, however, a larger sample group might have enabled any more subtle differences between either of the variations to become apparent in the results.

8.3 Discussion

NH1, the null hypothesis, cannot be rejected in terms of the usability criteria. There was no significant difference between the usability for Variations 1 and 2 from the t-test comparisons shown in Table 35, and so the alternative hypothesis is false. However, more errors were made on average on Variation 2 compared to Variation 1. This may partly be because it was easier to make errors more quickly using Variation 2, since a failed authentication attempt on the Variation 2 could be followed by a quicker attempt to re-order the chosen images, rather than having to go back to the beginning to choose all four, or five, images from scratch. This may also be the reason overall for the fact the average time taken for participants to log in using Variation 2 was shorter than that for Variation 1. From this NH1 can be partially rejected as there is some indication that there is a difference between the two variations.

Similarly to NH1, neither of NH2 or NH3 can be rejected in terms of the usability and security criteria. The results in Table 36 and Table 37 show that for none of the criteria was there a statistically significant difference in the usability scores between the original Awase-E model and either of the two variations. This also renders the alternative hypotheses, NHA2 and NHA3, false. However, in terms of the time taken and number of errors made, the null hypotheses NH2 and NH3 were shown to be false as there was a clear difference in the results for the two variations. For both variations, participants made more errors and took longer to use the applications than those participants using the original Awase-E method, as it shown in Table 38.

It is worth noting that seven participants using Variation 1 and five using Variation did not make any errors at all in logging in. This indicates that even with no training or prior

experience, some people would be able to use these variations. In combination with additional experience, it is expected that over time participants would not only become faster at logging in, but also make fewer errors once they became familiar with their own chosen images and the methods themselves.

Several participants also commented that they were choosing images according to a basic theme. One user for example specifically chose images related to food (since it was around lunch time and he was hungry). This generic theme of picking images often persisted across both variations. For example, three participants used a basic theme of animals (e.g. Figure 56) across both experiments, and several others used themes of food, or plants for each of the variations in Experiment 2.

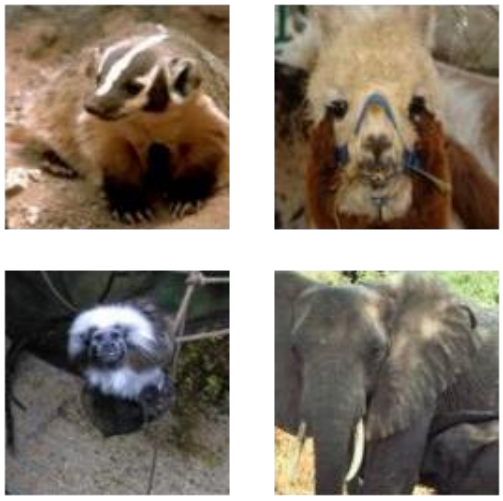


Figure 56: Example image set chosen for Variation 2 using an animal based theme

This planning of how to choose images is directly contrasted by other participants who did not seem to realise they would have to log in again using their chosen images, often through not reading the instructions properly or at all. This meant that they did not pay

much attention when choosing the original images, making the process of logging in using them later substantially more difficult.

It should also be noted that the memory load on participants was not just that of one set of password images, since each participant was being required to remember both sets of images at once. This would be a significant memory load over normal usage of the variations and so would have increased both the time taken and the number of errors, particularly if users became confused about which sets of images were from which authentication method. Care was taken to ensure that none of the images from one variation appeared in the other variation but there were some similar images that could be mistaken. This increase in error and time taken will have affected each of the variations equally though, and so does not affect the direct comparison between them.

This contradicts the work by Moncur & Leplâtre (2007) who showed that their participants were able to remember five different graphical passwords, whereas here they were only required to remember two. This may again be related to whether or not the participants were trying to remember them actively between uses as they might have been during Moncur's work.

Experiment 1 would have suffered less from this due to the different types of memory being tested. Each of the three original methods tested an entirely different type of memory, and the types of information were very separate, which would have allowed participants to more easily separate the different passwords in their heads. At least one of the participants from Experiment 2 was unable to log in using one of the variations due to using similar images on the other and getting confused between the two.

Overall the amount of time taken for users to log in is still very high. Even for the reduced set of people who did successfully login, the average time taken was 64 seconds and 50 seconds for Variation 1 and Variation 2 respectively. This is a very long period of time if a user wants access to something quickly. It is however expected that this time would significantly reduce as the user became accustomed to the login procedure and the images each had chosen as their pass images. Assuming the users were able to remember their images, and the order if necessary, it is expected that Variation 1 would be quicker to login, since it does not require the additional step of having to rearrange the images already chosen.

When compared to Experiment 1, there is a clear and significant drop in participant performances based on data from Dataset 2 relating to the time taken to perform each of the actions and the number of errors made. When compared to the original Awase-E method, the number of errors doubled for Variation 2, and increased a lot for Variation 1 (although it more than doubled when considering all participants rather than just those who successfully logged in). The time taken also increased significantly for the logging in part of the method, more than doubling for Variation 1 and increasing by 80% for Variation 2. Oddly though, the time taken to create an account did actually decrease when going from the original Awase-E to Variation 1, but it is likely this is a statistical blip rather than a trend, given Variation 1 is the same as the original Awase-E method, but with one additional image to choose.

An increase in the time taken by participants to perform the tasks in Experiment 2 (compared to Experiment 1) was expected for both of the new variations, along with an increase in the number of errors made. This is due to the increased memory load, under

which participants were being placed. A higher error rate would partially imply that the variations are less usable than the original.

Variation 2 can be shown to be more secure than Variation 1, since the chances of randomly guessing the answer for Variation 2 is 1 in 12^4 x 4! (= 497664) against 1 in 13^5 (= 371293) for Variation 1. This is purely in terms of the numerical probability, and there is the additional layer of security in that if an attacker does choose the wrong set of images, they would never be able to log in without being informed that the wrong images were being tried. Due to this, and the fact that people were often quicker at using Variation 2 rather than Variation 1, this would be a better variation to take forward as an improvement over the original Awase-E method.

As was the case with Experiment 1, scores given by the students were relatively high on average. None of the criteria has an average score across all participants lower than 5/7 which would be expected to indicate that the variations were both very usable. These results are likely to have been effected by the good subject effect (see Section 6.6.1.2) in the same way as Experiment 1, which would account for the discrepancy between users complaining they were unable to use the variations properly, and the scores given.

Having said that, the participants did seem able to understand how to use both of the variations, and the majority of issues were due to participants not being able to remember the images they had chosen. From this, the variations could indeed be said to be as usable as the analysed results from the questionnaires, along with the other information gathered, suggest.

9 Chapter Nine – Discussion

9.1 Research Objectives

The three research objectives were met as set out in Section 1.2. A set of criteria were created enabling the evaluation of graphical authentication methods. Recognition based methods were shown to be the most usable of the three types and the subsequent variations to one of the methods did increase the security without substantially negatively impacting the usability.

1. *Develop criteria to evaluate whether a graphical authentication method is usable and the most secure.*

Eleven criteria were formulated initially from a combination of sources, as detailed in Chapter Five. These criteria cover both usability and security, however the security aspects were exclusively objective ones and so were not evaluated during the experiment. These criteria were instead addressed in the implementation of the methods themselves. Two further criteria were added to the set to collect information about how the users of a system or method actually felt about it. This aspect is entirely subjective but many previous usability studies failed to look at how the methods or systems being evaluated were perceived by the users. The original six usability related criteria were not found to show adequately whether any particular method was usable, since in the first study, these criteria were all given high scores by the participants. But the feeling from many of the participants was that PassPoints was not usable at all. Two additional criteria relating to the user satisfaction element of usability highlighted more of the differences between the methods and so these two are justified in being included with the full set.

2. *Use the criteria to determine which type of graphical authentication method is the most usable.*

Three methods were evaluated, namely, Awase-E, DrawASecret and PassPoints. As outlined above, Awase-E was found to be the most usable from the results of Experiment 1 as was highlighted by the additional criteria and logging data. This was as anticipated since Awase-E relies on recognition memory rather than recall, which has been shown (from more general psychological experiments) to be easier for people to use. However, regardless of recognition memory, a lot of the participants seemed to understand the concept of the Awase-E method more intuitively than they did the other two methods, particularly PassPoints.

3. *Develop a new authentication method building on the results of the first experiment to account for any issues raised by the experiment relating to the usability or security of one or more of the methods*

Awase-E was also shown to be the least secure of the three methods evaluated in Experiment 1 since the password space for it was shown to be smaller than that of the other two. It is not actually possible to use a complex password with Awase-E, and this is an issue with the method which needs to be addressed in order to provide a more secure method which is still user friendly. Two variations were developed based on the original Awase-E method which were intended to make it more secure, without significantly impacting on its usability. The first of these involved using five images to remember as the password instead of four, from a set of over 100 images, for the user to remember as their password. The second required the user to remember just the original number of four images and put them into an order, so when logging in they had to remember both the set of images and the order they were in. Experiment 2 was run using a different set of student

participants from those used in Experiment 1 in order to see which of the two variations was preferred.

Although Awase-E was the least secure of the methods, it was shown to be the most usable of the three. Both of the variations created would improve the security of Awase-E, and so the second experiment was required to show both which of the two methods was the most usable and to confirm that the usability of this preferred variation was not significantly reduced over that of the original method. By doing so, the final research objective was met.

9.2 Summary of thesis

Three separate empirical studies into the field of mobile security and usability were performed as part of this thesis. The first was a Systematic Mapping Study of the general area of mobile security. This produced some interesting results about where the majority of the work is being performed within the field, showing that a lot of work is concentrated in the area of authentication. However, it also highlighted that there was a lack of research discussing the usability aspect of many of the new systems or designs which were being produced.

There were issues with the scope of the Mapping Study though which are likely to have affected the results. The most important of these is the limitation to 150 papers from each of the digital libraries, and so to 450 papers overall. This resulted in some relevant papers being omitted from the mapping study. Despite this, there was still considered to be a lack

of research on usability within the field. Hence a further literature review concentrating on the authentication element was carried out.

Authentication (or Identity Establishment) was shown to be the most popular area of work within the field of Mobile Security, and is one of the first interfaces a user can expect to use with any system. The usability of a system at this point is vitally important in ensuring the process for a user is as smooth as possible. The papers found from the Systematic Mapping Study provided a good starting point for the subsequent literature review into authentication, and more specifically mobile authentication, since it also showed where work was being published in this area.

This subsequent study showed that there are a significant number of alternative authentication methods to the standard alphanumeric username and password combination used in a significant number of systems today. One of the most promising of these is graphical authentication methods, which replaces alphanumeric strings with some form of image (be it a picture or drawn) which the user has to remember instead. Various pieces of psychological research have shown that images are easier than strings of text to remember (Shepard, 1967, Tulving & Watkins, 1973, Hollingworth, 1913), so would allow for a more complicated password to be used whilst making it easier to remember. A graphical authentication method would in theory be relatively easy to implement on a mobile device, since they require no additional hardware to be fitted (such as a fingerprint scanner or better quality camera).

Graphical authentication methods can be broken down into three types, recognition based, recall based and cued recall based. Again it was found that there was a lack of research on

the usability of the types of authentication being performed, particularly in terms of comparative studies of different types of methods. Several of the published studies tested a new method against an original method to show the differences and improvements which may have been added, however there was no works comparing methods of different types.

A further literature review was performed into the area of usability to find guidelines and rules with which to evaluate some of the graphical authentication methods described in Chapter Three. The full results of this are discussed in Chapter Six, however perhaps the most important point is that it has been shown that tasks based on recognition memory ought to be easier for people to perform than those based on full recall. Relating this to authentication this means that authentication methods where the user only has to recognise a series of objects ought to be easier for people to use than one where they are required to remember the whole object or objects from memory.

In order to evaluate the different types of graphical authentication in terms of their usability, a set of criteria was developed using both usability guidelines and mobile security guidelines. Eleven criteria were generated in this way, some relating to usability and some relating to security. The usability related criteria are subjective measures and so cannot be evaluated within the use of an empirical experiment to determine people's opinions of the methods. These criteria were extended based on the fact that they did not adequately involve the user satisfaction element of usability, originally concentrating on effectiveness and efficiency, thus bringing the number of criteria to a total of 13.

The lack of criteria relating to user satisfaction does perhaps highlight an issue with some of the existing usability guidelines in that many do not consider the user satisfaction aspect

as strongly as the other two aspects. These three aspects are the basis of usability for any system and ought to be considered equally. If nothing else, it has been shown that when users do not like a security method, or find it frustrating, they will often find ways to circumvent the security elements of the system or disable them altogether (Adams & Sasse, 1999, Balfanz et al., 2004, Karlog et al., 2004, Dingledine & Mathewson, 2006). The development of these criteria to evaluation the usability of authentication methods on mobile devices was the first of the key research objectives of the thesis.

Experiment 1 was designed to compare the usability of different types of authentication method. Three different methods were chosen for to be compared in Experiment 1 with the strategy for each method representing one of the three different types of authentication. Awase-E is a recognition method involving remembering pictures, DrawASecret is a recall based method based on drawing a pattern on the screen, and PassPoints is a cued recall method based on remembering areas on a picture.

Three hypotheses were proposed for Experiment 1. H01 hypothesised that there was no difference between the methods and this was shown to be true for some of the criteria and not for others. Most notably it was rejected for the criteria relating to how the user experienced the different methods as a whole, rather than specific aspects of the method (C13). This showed that people did feel the methods were different in terms of usability, even though they all did the same job.

H02 hypothesised that the usability scores between participants would be no different when the users had different amounts of previous experience of mobile devices. It was expected that there would be a significant difference between those who were

inexperienced with mobile devices and those who were more familiar with them, as the former would find it more difficult to cope with the methods, and so enabling H02 to be rejected. Unexpectedly, it was shown that almost universally, the average scores from Group A and Group B were not statistically significantly different. Additionally, there was shown to be no significant difference within the groups when comparing the scores given by participants who own, and those who do not own, smartphones already. This is surprising since it was expected that participants who were used to the technology would find the applications more familiar and so easier to use, meaning they scored higher. The only place where this familiarity seemed to make a difference was that the time it took the participants to use each method did vary, with participants from Group A taking a much shorter time on average than those in Group B, who were less experienced with mobile technology.

H03 related to the effects of periodicity, or order, on the usability scores. Period effects within an experiment would mean that a method tested first or third would perform better than if it was tested second. This effect was not shown to be present within the data, possibly because it was masked by the stronger effects of lack of experience of some of the participants.

Although some of the results from Experiment 1 were not particularly indicative of which of the three methods was the most usable, criteria related to user satisfaction along with the qualitative data gathered did show that Awase-E was the preferred choice of method. However it was also the least secure of the three methods, having a relatively small password space (particularly when compared to the limitless password space of

DrawASecret). It was therefore decided to extend this method in order to make it more secure but without having a significant impact on the usability.

Two new variations were proposed for the Awase-E method, the first based on increasing the number of images the user has to remember (Variation 1), and the second requiring the order of the original four images to be remembered in addition to the images themselves (Variation 2). Both of these improved the security of the method by improving the potential password space available, and in the case of the Variation 2, adding an additional step altogether to the process. These two variations are discussed in more detail in Section 7.2. In Experiment 2, the setup was used to evaluate the two variations. This allowed for a direct comparison between Variation 1, Variation 2, and the original Awase-E method.

Two hypotheses were proposed for Experiment 2, the first (NH1) dealing with the difference between the Variation 1 and 2. It was shown that Variation 2 (the order variation) was marginally more usable than Variation 1 (the additional images variation) based on the qualitative data gathered, although there was again no statistically significant difference in the usability scores. However, the time taken by participants and the number of errors made confirms that the order variation was more usable.

The second and third hypotheses (NH2 and NH3) looked at the relationship between the two variations and the original Awase-E method. There was shown to be no statistically significant difference between either of the two new variations and the original Awase-E method, based on the usability criteria. This is a good sign for improving the security of methods since there is normally a balance between the usability and security of any system, where increasing one often decreases the other. Although the results were not

statistically significant, there were indications that the variations of Awase-E evaluated in Experiment 2 across all of the criteria on average scored lower than the realisation of Awase-E evaluated in Experiment 1, and this is supported by some of the logging data. The number of errors made by participants was on average much larger for both Variations 1 and 2 in Experiment 2 than for Awase-E in Experiment 1 along with the time taken to log in being longer. So the number of login attempts, and time taken, increased with both the increase in the number of images to be remembered and the addition of the order. There was no statistically significant difference between the Variations 1 and 2 based on the usability criteria.

Overall, this shows that simply adding more information to the password decreased the usability more than adding context to the information already there, specifically an order for the images the user has to remember. Even in terms of normal passwords, this shows that asking users to remember longer passwords is less effective than allowing them to keep shorter ones, but requiring some sort of contextual information about the password. From a memory standpoint, this would suggest that recognising four images in order is easier than recognising five separate images.

9.3 Lessons learnt

Clearly the number of papers included in the Mapping Study was a major limitation on the results and conclusions arising from it, and the most obvious way to improve this would have been to increase the number of papers included. The search string itself could also have been further refined to reduce the number of irrelevant papers returned by the digital libraries. From the 450 total papers found by the queries, only 214 were finally included. These were spread reasonably evenly through the results set so even by the end there were

papers being included rather than the number of included papers trailing off as might be expected. This however was always likely to be a side effect of the fact that the Mapping Study concentrated on such a wide area (mobile security in general rather than concentrating on mobile authentication for example).

In terms of the two experiments, one of the main issues faced by many participants was a lack of understanding of how the implementations of the methods worked, or what they as participants were expected to do. This was partially intentional in that one of the outcomes was to see how well participants did cope with something unfamiliar to them. However, it made many participants feel uncomfortable to have to ask for help.

More detailed instructions presented on screen would confuse users even more since the size of other objects on the screen at any given time would have to be reduced to compensate for the extra text being added. Similarly it was shown that many of the users did not thoroughly read the instructions that were provided and had to ask questions which would not have been necessary had taken in the information on the instructions screen show at the beginning.

Many of the limitations which applied to Experiment 1 also apply to Experiment 2 (see Sections 5.2.5 and 6.6). It was decided to keep as many of the variables as possible the same to allow for a better comparison between the two variations and the original Awase-E method. The same number of student participants were used, 18, which could be increased if a more in depth study was undertaken. Although Experiment 1 found no difference between the two original groups, the selection process was still biased and so would benefit from a random participant selection process.

In Experiment 2, the create account stage of Awase-E was not modified to allow for a confirmation of the images chosen by the user. This was done to ensure participants using the variations did not have an advantage over those using the original method in Experiment 1. For both Experiments therefore, when logging in, the participants had only seen the images they chose once before, rather than confirming and possibly cementing the images more into their memories. This would have perhaps helped several of the participants who were under the impression they would not have to remember the images, despite being informed of this before starting the experiment.

The lack of confirmation added to the effect of one of the questions on the questionnaire not being applicable to all users if they made no mistakes when using the method, since they were given fewer tasks to perform and so fewer changes to make mistakes, as discussed in Section 6.3.2.

The criteria by themselves were shown in Experiment 1 to be inadequate as discriminators of usability for the different methods. The additional C12 and C13 criteria were able to highlight some of the differences suggested from the qualitative data collected during both of the experiments. The questions relating to these criteria were included for both experiments as a more direct method of finding the participants' opinions on the usability of the different methods. These should be included with the criteria for any future experiments to allow participants to convey their opinions of the methods being investigated.

The work has also highlighted some of the psychological issues associated with performing experiments. Many of the scores given by participants were relatively high and were not

supported by the comments made verbally during the actual experiments. The scores given by the participants rarely averaged less than 4 out of a maximum of 7 despite many of the participants expressing their dislike of one or more of the methods. The most likely reason for this is the participants attempting to either give the 'right' answers to make themselves look better, or to give the answers they thought the researcher would want. As mentioned in Chapter Six, it is difficult to avoid this issue entirely.

The number of participants taking part in both experiments was relatively low. This means that the statistical power of the experiments is not as strong as it perhaps could have been if more participants were used and any further studies would benefit from a significantly increased sample size.

The sample itself is not a random sample of the entire population and so a more random sample would produce more generalisable results. As it stands, the results from Experiment 1 can only be applied to the two subsets of the population which were used as participants, and the results from Experiment 2 only to one subset of the population i.e. computing students. Other groups of people may find the usability of the different applications significantly difference due to their differing needs from a security method (business users may required a higher level of security if they store confidential information on the device). Neither the student group from both experiments or Group B from Experiment 1 are likely to have highly confidential information stored on their phone, and so would be less inclined to require a particularly secure authentication method to gain access. Businesses however may want more security to protect more sensitive information, and so be inclined to overlook some of the usability issues related to the

DrawASecret and PassPoints methods in favour of the higher security they offer, particularly if the device is a word related one rather than personal.

One other aspect not addressed in the design of Experiments 1 and 2, is the effect of learning on the outcomes. Each of the participants had only one set of attempts at creating a password and logging in using each of the methods. This is not comparable to a real life situation where a user would be performing authentication on a regular basis. How often the user authenticated themselves on the device would be dependent on the specific implementation and the data it was required to protect, but any regular use of the method would still be more than the one off situation experienced during the experiment.

A delay between the creation of a password and logging in was factored into the design of the experiment, however this was only between 15 and 20 minutes. A distraction task was deliberately introduced to try to ensure the participant did not spend the time repeating the password over and over mentally to remember it, so as to simulate a longer time between the creation and login. However again, this was still a one off process with no real opportunity for learning the password properly.

A longer term study could be carried out which would evaluate the user's ability to remember the password over many days or weeks. If the user was required to remember the password many times over a set time period then it is expected that as the duration of the experiment increased, the number of errors made by users would decrease towards zero due to the effect of learning. What effect this has on the usability would need to be investigated. It might be expected that the less usable methods in the short term would become more usable as time passed and people became more familiar with the concept.

On a similar note, there was no element of training within the experimental setup. Training would involve giving participants an example of how the method was should be used before allowing them to try it themselves. In a number of cases this would have helped participants in both Experiments 1 and 2 since it was clear that some did not understand the concepts behind the methods and so chose poor or unmemorable passwords (e.g. Figure 35). Again, what effect this would have on the usability scores is not known, but it is likely to increase them since participants are less likely to get frustrated through a lack of understanding and score the method poorly for usability.

Another possible way to investigate the effects of learning on the outcomes of the experiments would be to require the participants to see a trial run through each of the methods. This could be performed by the researcher to show participants how the method could, or ought to, be used. Similarly, there could be a separate part for each of method specifically for training users with extra instructions on the screen to allow for a practice run. In the case of Awase-E and its extensions this would have to have an additional set of images which were different to those in use for the actual method.

As a final point, the methods themselves can currently only be used as a locking method for specific applications on a user's device. For the method to be used as the unlock code for the phone as a whole, Google, as android's author, would need to include it in the operating system itself. Given the recent push towards facial recognition in the latest version of android, this seems unlikely to be achievable. However, it would be possible for the method to be used to lock specific applications on the phone from unauthorised access, as is done with a variety of other security applications available already.

10 Chapter Ten – Conclusions

This final chapter highlights some of the potential further research which could be performed building on the findings of the research reported here. It also provides some concluding remarks on the results of the research as a whole.

10.1 Concluding Remarks

This thesis has highlighted the differences between recall based and recognition based methods of authentication, specifically relating to graphical authentication. Of the three methods initially evaluated in Experiment 1, Awase-E was found to be the most usable although it was not the most secure. One of the more secure methods was also shown to be the least popular, and so the least likely to be actually used on a day to day basis. Two improved variations of the Awase-E method were implemented and evaluated, one using additional images and the other requiring the images to be ordered. Variation 2 (ordering the images) was shown to be the most usable from Experiment 2.

Through performing the experiments required to evaluate the methods, the three main research objectives were met. A set of criteria were developed against which the methods were evaluated in terms of their security and usability. There is room for improvement, or expansion, of the criteria in order to factor in further requirements from a user's perspective relating to an authentication system, since not all of the participants comments from the experiments were reflected in the scores given to the methods by the participants.

More generally some of the main issues related to recognition based graphical authentication methods were also highlighted, the most prominent being the reduced

password space compared to that of the recall or cued recall based methods. Since recognition methods always need to prompt the user with a set of information for the user to recognise, this deficit will always be a disadvantage over the other methods, such as DrawASecret, which has an almost infinite theoretical password space. Unfortunately, along with the increased password space associated with DrawASecret comes a decrease in the ability of users to fully recall their password accurately.

To achieve a very secure password to protect a mobile device, the password must be one which a user is less likely to be able to reproduce each time they need to. The goal is to create a password which is memorable enough for a user to reproduce accurately and quickly even whilst distracted by other real world events. The password must also be resistant to shoulder surfing style attacks as well as not being so obvious that an attacker could easily guess it. Additionally, the method by which the password is entered needs to be designed to be simple enough for both novice and less technologically capable users to be able to use.

Experiment 1 showed that of the three main types of graphical authentication method, recall, recognition and cued recall, that recognition methods were the most usable of the three.

Both Experiment 1 and Experiment 2 were able to show some differences between the methods being evaluated, however in both cases there were clearly other factors influencing the results and perhaps masking some of the subtler differences between the methods. Improving on the design of the experiments, perhaps using more participants, or retesting the participants after a more significant time period had elapsed would improve

the quality of the results as well as perhaps highlight the differences between the methods more clearly. In the case of Experiment 1, it is unlikely that the outcome would change in terms of which of the methods the participants found the most usable, based on the qualitative data collected during the experiment. With Experiment 2, there was not much difference between the two methods, so a longer study, or one with more participants to increase the statistical power of the results, could provide different results.

The Awase-E method, with the additional modification of the order of the images adding to the complexity of the password, is a step towards creating a method which is both usable and secure. Although it was shown from the experiments that the participants spent some time both creating their password and logging in, it is felt this would decrease significantly as the user became more familiar with the method. The creation of the password is a onetime only step, and once the user was used to logging in, it would be unlikely to take them more than a few seconds to choose the correct images and reorder them. A modified Awase-E method, such as Variation 2, meets the original research objectives set out in Section 1.2.

10.2 Further Research

The work done in this thesis could be extended into other areas of graphical authentication. Other promising graphical methods are available, both for mobile and non mobile devices, some of which were described in Chapter Three. Many of these have not been subjected to a usability evaluation, nor directly compared with the other available methods.

Work still needs to be done to try to find a way to create a graphical authentication method with the increased security benefits of PassPoints or DrawASecret, but without the loss of usability associated with them. How feasible this would be to achieve whilst using only recognition memory methods would also need investigating.

The criteria developed could be extended further to include more assessment of the security of the methods being evaluated. All of the methods used in the research reported in this thesis met the security related criteria formulated from published criteria to mobile security. Despite this, there was a significant different in the level of security offered by the different methods, with Awase-E being the least secure of the three original ones. The criteria could be extended to include a requirement for a certain level of complexity of password being stored on the device, and how many such passwords would be feasible to be used by a human. For example, DrawASecret technically has an unlimited password space, however after a certain length it is unlikely any passwords created would be memorable enough to be recalled accurately even though a pattern was used rather than an alpha numeric string.

An empirical study would need to be undertaken to see at what level the complexity of a drawn graphical password becomes too high and so make an estimate of the real password space for each type of authentication method. This could be extended to both the recognition and cued recall methods to determine at what level of complexity of passwords the methods become too unusable. In the case of PassPoints this would be in the form of changing the size of the grids used. Making the individual squares larger would make the method easier to use since the error margin would be larger when logging in, however it

would also be less secure both because of the increased error margin and the lower number of possible areas on the image to choose from.

Another possible extension would be to look into some of the methods which are meant to counter shoulder surfing styles of attack. Only PassPoints has some protection from this style of attack due to the nature of choosing a point in a generic area, but even here it would be fairly obvious which areas on the screen the user was choosing. Finding a way to incorporate some of the pre existing counter measures against shoulder surfing into any of the methods presented would significantly increase their security. How much this affected the usability of the methods would have to be investigated as well.

Another interesting topic would be to further study the differences in usability between people with different levels of computing literacy or technology experience. It seems reasonable that those more familiar with a technology will be able to use it more effectively, however, Experiment 1 does not support this assumption. This could of course mean that the methods are sufficiently usable for the majority of people and hence no further usability testing would be needed.

No peer reviewed academic surveys have been found into what percentage of the population, be they business or home users, regularly use any form of security method or protection on their phones. The two surveys that were found were both commissioned by companies who specialised in producing alternative authentication methods for devices, and so would have had a vested interest in showing that people are not secure enough with their phones.

It is anecdotally considered that a large proportion of people do not use any authentication on their mobile device, meaning that if the device is lost, or stolen, then whoever ends up with it can easily just look through anything stored on the device. The reasons often cited by people for not using authentication methods are a response to the time taken to use it and other usability concerns. The work in this thesis is intended to make a contribution to addressing these concerns. A full study in this area would highlight the actual risks being taken both by companies and individuals in terms of their security practices and would provide a basis for trying to ensure people are more inclined to use the authentication methods available. This applies not only to authentication methods, but also on a wider scale to security methods as a whole. No matter how secure a security method is, if it is considered unusable then there is no real value to it.

Glossary

Access control: Once the user has authenticated, access control is the process which limits that user to only the resources they are meant to be able to use within the system. This is often referred to as authorisation.

Authentication: The process whereby a user confirms their identity to a system by passing over some form of details which only they would know, have or is a characteristic of them (e.g. a fingerprint). This is a synonym of identity establishment.

Availability: Ensuring that a resource, be it a device on the network or the network itself, is accessible by preventing or minimising the effect of denial of server style attacks.

Awase-E: One of the strategies for authenticating a user on a mobile device through the use of pictures and based on recognition only. This is referred to as the Awase-E method.

Data and Message Security: Protection of data stored both on the device and transmitted to and from it. This is usually performed through some form of encryption process to prevent an eavesdropper from reading it directly.

Dataset 1: Scores for each question in the questionnaire used in each of the experiments.

Dataset 2: Logging data recorded on the device during the experimental process with each participant. This includes the patterns drawn by each participant, the passwords created or images chosen as well as each of the button presses made by the participant. Also recorded is a timestamp of each action.

Dataset 3: Qualitative data from the participants collected during each experiment such as comments made about the methods and the general approach of the participant towards the experiment.

DrawASecret: One of the strategies for authenticating a user on a mobile device through the use of drawings and based on full recall. This is referred to as the DrawASecret method.

Experiment 1: An experiment performed to evaluate the differences in usability between three different graphical authentication methods, Awase-E, DrawASecret and PassPoints.

Experiment 2: An experiment performed to evaluate the differences in usability between two variations of the Awase-E graphical authentication method. The first involving an increased number of pass images, referred to as Variation 1, and the second requiring the order of the images chosen to be remembered, referred to as Variation 2.

Group A: The group of participants in Experiment 1 selected by asking for volunteers from the undergraduate population studying computing courses at Keele University.

Group B: The group of participants in Experiment 1 selected using a convenience sample of the general population.

Identity Establishment: The process whereby a user confirms their identity to a system by passing over some form of details which only they would know, have or is a property of them. This is a synonym of authentication.

Implementation (of a method): The specific implementation used in the experiments to evaluate the methods in question, be they the original method or the variations on it.

Method: A method of performing authentication based on a particular strategy (e.g. Awase-E is an authentication method using the strategy of using images as the user's password).

Method Effect: The effect on the usability scores from the questionnaire data which was caused by the differences in the methods used.

Non-repudiation: Preventing a user or device from falsely denying its part in an activity. This is often done through the use of an audit trail to record the actions of individual users.

PassPoints: One of the strategies for authenticating a user on a mobile device through the use of areas on a picture, and is based on cued recall. This is referred to as the PassPoints method.

Security Suite: A system which incorporates all of the security elements

Strategy: A generic way of authenticating a user, such as through the use of images.

Variation 1: A variation of the original Awase-E authentication method requiring an additional image to be remembered as part of the password increasing the number of images required from four to five.

Variation 2: A variation of the original Awase-E authentication method requiring the order of the four images picked to be remembered as part of the password.

References

Abadi, M. & Needham, R. (1996). Prudent Engineering Practice for Cryptographic Protocols. *IEEE Transactions on Software Engineering.* 22 (1): 6-15.

Adams, A. & Sasse, M.A. (1999). Users are not the enemy. *Communications of the ACM.* 42 (12): 40-46.

Adams, C., Farrell, S., Kause, T. & Mononen, T. (2005). *RFC 4210 - Internet X.509 Public Key Infrastructure Certificate Management Protocol (CMP).* Available from: http://tools.ietf.org/html/rfc4210.

Ahituv, N., Lapid, Y. & Neumann, S. (1987). Verifying the authentication of an information system user. *Computers & Security.* 6 (2): 152-157.

Almenárez, F. & Campo, C. (2003). SPDP: A Secure Service Discovery Protocol for Ad-hoc Networks. In: *Workshop on Next Generation Networks.* Springer Verlag. pp. 9-9.

Al-muhtadi, J., Mickunas, D., Campbell, R. & Ranganathan, A. (2002). A flexible, privacy-preserving authentication framework for ubiquitous computing environments. In: *Proceedings of the 22nd International Conference on Distributed Computing Systems Workshops.* Vienna, Austria, July 2-5, Vienna, Austria: IEEE Computer Society. pp. 771-776.

Al-Qayedi, A., Adi, W., Zahro, A. & Mabrouk, A. (2004). Combined Web/mobile authentication for secure Web access control. In: *IEEE Wireless Communications and Networking Conference.* Atlanta, Georgia, 21-25 March, IEEE Computer Society. pp. 677-681.

Alsos, O.A. & Dahl, Y. (2008). Toward a best practice for laboratory-based usability evaluations of mobile ICT for hospitals. In: *Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges.* NordiCHI '08. Lund, Sweden, New York, NY, USA: ACM. pp. 3-12.

Android Protector *Android Protector.* Available from: http://www.android-password.com/. [Accessed: 02/03/2012].

Animetrics, I. (2011). *FaceR Credential Me.* Available from: http://www.animetrics.com/products/CredentialMe.php. [Accessed: 15/08/2011].

Arora, D., Raghunathan, A., Ravi, S., Sankaradass, M., Jha, N.K. & Chakradhar, S.T. (2007). Exploring Software Partitions for Fast Security Processing on a Multiprocessor Mobile SoC. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems.* 15 (6): 699-710.

Aviv, A.J., Gibson, K., Mossop, E., Blaze, M. & Smith, J.M. (2010). Smudge attacks on smartphone touch screens. In: *USENIX 4th Workshop on Offensive Technologies.*

Bae, I., Lee, H. & Lee, K. (2007). Design and evaluation of a rough set-based anomaly detection scheme considering weighted feature values. *International Journal of Knowledge-Based and Intelligent Engineering Systems.* 11 (4): 201-206.

Balagtas-Fernandez, F. & Hussmann, H. (2009). A Methodology and Framework to Simplify Usability Analysis of Mobile Applications. In: *IEEE/ACM International Conference on Automated Software Engineering.* ASE '09. Washington, DC, USA, IEEE Computer Society. pp. 520-524.

Balfanz, D., Durfee, G., Grinter, R.E. & Smetters, D.K. (2004). In Search of Usable Security: Five Lessons from the Field. *IEEE Security and Privacy.* 2 (5): 19-24.

Bardram, E. (2005). The trouble with login: on usability and computer security in ubiquitous computing. *Personal Ubiquitous Computing.* 9 (6): 357-367.

Barton, B.F. & Barton, M.S. (1984). User-friendly password methods for computer-mediated information systems. *Computers and Security.* 3 (3): 186-195.

Benantar, M. (2006). *Access Control Systems: Security, Identity Management and Trust Models.* USA: Springer.

Ben-Asher, N., Meyer, J., Moller, S. & Englert, R. (2009). An Experimental System for Studying the Tradeoff between Usability and Security. In: *International Conference on Availability, Reliability and Security.* Los Alamitos, CA, USA: IEEE Computer Society. pp. 882-887.

Bevan, N. (1995). Measuring usability as quality of use. *Software Quality Journal.* 4 (2): 115-130.

Bevan, N. (1995). Usability is Quality of Use. In: *Proceedings of the Sixth International Conference on Human-Computer Interaction.* Elsevier. pp. 354.

Bevan, N. & Macleod, M. (1994). Usability measurement in context. *Behaviour and Information Technology.* 13 (1-2): 132-145.

Bharghavan, V. (1994). Secure Wireless LANs. In: *Proceedings of the 2nd ACM Conference on Computer and communications security.* Fairfax, Virginia, USA, 2-4 November, ACM. pp. 10-17.

Bharghavan, V. & Ramamoorthy, C.V. (1995). Security issues in mobile communications. In: *Proceedings of the Second International Symposium on Autonomous Decentralized Systems.* Toronto, USA, 25-27 April, Washington, DC, USA: IEEE Computer Society. pp. 19-24.

Bhattacharyya, D., Ranjan, R., Das, P., Tai-hoon Kim & Bandyopadhyay, S.K. (2009). Biometric Authentication Techniques and its Future Possibilities. In: *Second International Conference on Computer and Electrical Engineering.* Dubai, 28-30 December, IEEE. pp. 652-655.

Biddle, R., Chiasson, S. & van Oorschot, P.C. (2011). *Graphical Passwords: Learning from the First Twelve Years.* TR-11-01. School of Computer Science: Carleton University.

Bin Nafey, F. & Ramanaiah, O.B.V. (2008). A study on Rijndael algorithm for providing confidentiality to mobile devices. In: *TENCON, IEEE Region 10 Conference.* Hyderabad, India, 20-23 November, IEEE Computer Society. pp. 1-6.

Birget, J., Hong, D. & Memon, N. (2003). *Robust Discretization, With an Application to Graphical Passwords.* 168. Cryptology ePrint Archive.

Blonder, G.E. (1996). *Graphical Passwords.* vol. 5,559,961. US: .

Bonneau, J., Preibusch, S. & Anderson, R. (2012). A birthday present every eleven wallets? The security of customer-chosen banking PINs. In: *Financial Cryptography and Data Security 2012.* 27 February - 2 March, .

Botha, R.A., Furnell, S.M. & Clarke, N.L. (2009). From desktop to mobile: Examining the security experience. *Computers & Security.* 28 (3-4): 130-137.

Braz, C. & Robert, J. (2006). Security and usability: the case of the user authentication methods. In: *Proceedings of the 18th International Conference of the Association Francophone d'Interaction Homme-Machine.* IHM '06. Montreal, Canada, New York, NY, USA: ACM. pp. 199-203.

Brewster, S. (2002). Overcoming the Lack of Screen Space on Mobile Computers. *Personal Ubiquitous Computing.* 6 (3): 188-205.

Brostoff, S. & Sasse, M.A. (2000). Are Passfaces more usable than passwords? A field trial investigation. In: *Proceedings of Conference on Human-Computer Interaction 2000.* Sunderland University, England, pp. 424.

Brown, A.S., Bracken, E., Zoccoli, S. & Douglas, K. (2004). Generating and remembering passwords. *Applied Cognitive Psychology.* 18 (6): 641-651.

Buchanan, G., Farrant, S., Jones, M., Thimbleby, H., Marsden, G. & Pazzani, M. (2001). Improving Mobile Internet Usability. In: *Proceedings of the 10th International Conference on the World Wide Web.* WWW '01. Hong Kong, New York, NY, USA: ACM. pp. 673-680.

Budgen, D. (2008). *ASE-EVAL: Software Evaluation Glossary – Version 1.1.* Available from: http://www.dur.ac.uk/ebse/resources/ASE-EVAL-glossary.pdf. [Accessed: 07/07/2010].

Buennemeyer, T.K., Nelson, T.M., Clagett, L.M., Dunning, J.P., Marchany, R.C. & Tront, J.G. (2008). Mobile Device Profiling and Intrusion Detection Using Smart Batteries. In: *41st Annual Hawaii International Conference on System Sciences.* Waikoloa, Big Island, Hawaii, January 7-10, IEEE Computer Society. pp. 296-296.

Burr, W.E., Dodson, D.F. & Polk, W.T. (2006). *Electronic Authentication Guideline.* Available from: http://csrc.nist.gov/publications/nistpubs/800-63/SP800-63V1_0_2.pdf. [Accessed: 08/01/2010].

Calin, B. (2009). *Statistics from 10,000 leaked Hotmail passwords.* Available from: http://www.acunetix.com/blog/news/statistics-from-10000-leaked-hotmail-passwords/. [Accessed: 10/11/2011].

Carey, M.J. (2012). *Rapid7 SecurityStreet: It's Time to Ban Bad Passwords.* [Online] 11/06/2012. Available from: https://community.rapid7.com/community/infosec/blog/2012/06/11/its-time-to-ban-bad-passwords. [Accessed: 13/06/2012].

CarrotApp *App Protector Pro.* Available from: http://www.carrotapp.com/portfolio_1/app-protector/. [Accessed: 02/03/2012].

Carvalho, M. (2008). Security in Mobile Ad Hoc Networks. *IEEE Security and Privacy.* 6 (2): 72-75.

Chen, H. & Sivakumar, T.V.L.N. (2005). Access control for future mobile devices. In: *IEEE Wireless Communications and Networking Conference.* New Orleans, LA, USA, 13-17 March, IEEE Computer Society. pp. 1527-1532.

Chen, H. & Sivakumar, T.V.L.N. (2005). New authentication method for mobile centric communications. In: *61st IEEE Vehicular Technology Conference.* Stockholm, Sweden, May 30 - June 1, IEEE Computer Society. pp. 2780-2784.

Cheng, A., Wu, C., Ho, J. & Lee, D.T. (2004). Secure transparent mobile IP for intelligent transportation systems. In: *IEEE International Conference on Networking, Sensing and Control.* Taipei, Taiwan, 21-23 March, IEEE Computer Society. pp. 495-500.

Chiasson, S., Biddle, R. & van Oorschot, P.C. (2007). A second look at the usability of click-based graphical passwords. In: *Proceedings of the 3rd symposium on Usable privacy and security.* SOUPS '07. Pittsburgh, Pennsylvania, New York, NY, USA: ACM. pp. 1-12.

Chiasson, S., Forget, A., Biddle, R. & van Oorschot, P.C. (2009). User interface design affects security: patterns in click-based graphical passwords. *International Journal of Information Security.* 8 (6): 387-398.

Chiasson, S., Forget, A., Biddle, R. & van Oorschot, P.C. (2008). Influencing users towards better passwords: persuasive cued click-points. In: *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction - Volume 1.* BCS-HCI '08. Liverpool, United Kingdom, Swinton, UK, UK: British Computer Society. pp. 121-130.

Cho, Y.S. & Bao, L. (2006). Secure Access Control for Location-Based Applications in WLAN Systems. In: *IEEE International Conference on Mobile Adhoc and Sensor Systems.* MASS. Vancouver, Canada, 9-12 October, IEEE Computer Society. pp. 852-857.

Chowdhury, S. & Poet, R. (2011). Comparing the usability of Doodle and Nikon images to be used as authenticators in graphical authentication systems. In: *2011 International Conference on User Science and Engineering.* i-USEr 2011. Shah Alam, Selangor, 29 Nov - 01 Dec, pp. 54-58.

Confident Technologies (2011). *Survey Shows Smartphone Users Choose Convenience Over Security.* Available from: http://www.confidenttechnologies.com/news_events/survey-shows-smartphone-users-choose-convenience-over-security. [Accessed: 23/03/2012].

Coulouris, G., Dollimore, J., Kindberg, T. & Blair, G. (2012). *Distributed Systems: Concepts and Design.* 5th edn. Pearson Education Inc.

Dahl, Y., Alsos, O.A. & Svanæs, D. (2010). Fidelity Considerations for Simulation-Based Usability Assessments of Mobile ICT for Hospitals. *International Journal of Human-Computer Interaction.* 26 (5): 445-476.

de Kock, E., van Biljon, J. & Pretorius, M. (2009). Usability evaluation methods: mind the gaps. In: *Proceedings of the 2009 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists.* SAICSIT '09. Vanderbijlpark, Emfuleni, South Africa, New York, NY, USA: ACM. pp. 122-131.

De Luca, A., Denzel, M. & Hussmann, H. (2009). Look into my eyes!: Can You Guess My Password? In: *Proceedings of the 5th Symposium on Usable Privacy and Security.* SOUPS '09. Mountain View, California, New York, NY, USA: ACM. pp. 7:1-7:12.

Decker, C., Nguissi, S., Haller, J. & Kilian-Kehr, R. (2004). Proximity as a security property in a mobile enterprise application context. In: *Proceedings of the 37th Annual Hawaii International Conference on System Sciences.* Big Island, HI, USA, 5-8 January, IEEE Computer Society. pp. 10.

Deese, J. & Kaufman, R.A. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of experimental psychology.* 54 (3): 180-187.

Derawi, M.O., Nickel, C., Bours, P. & Busch, C. (2010). Unobtrusive User-Authentication on Mobile Phones Using Biometric Gait Recognition. In: *Sixth International Conference on Intelligent Information hiding and Multimedia Signal Processing.* IIH-MSP. Darmstadt, 15-17 October, pp. 306-311.

Dhamija, R. & Perrig, A. (2000). Déjá Vu - A User Study Using Images for Authentication. In: *SSYM'00: Proceedings of the 9th conference on USENIX Security Symposium.* Denver, Colorado, Berkeley, CA, USA: USENIX Association. pp. 4-4.

Dietiker, K. (2008). PGP whole disk encryption: blazing trails in IT security at UW Medicine. In: *Proceedings of the 36th annual ACM SIGUCCS fall conference.* Portland, Oregon, USA, 19-22 October, New York, NY, USA: ACM. pp. 17-20.

Dingledine, R. & Mathewson, N. (2006). Anonymity loves company: Usability and the network effect. In: *Fifth Workshop on the Economics of Information Security.* Sixth Workshop on Privacy Enhancing Technologies. Cambridge, UK, June, IEEE.

Doja, M.N. & Kumar, N. (2008). Image Authentication Schemes against Key-Logger Spyware. In: *Ninth ACIS International Conference on Software Engineerng, Artificial Intelligence, Networking and Parallel/Dsitributed Computing.* Phuket, 06-08 August 2008, pp. 574-579.

Duh, H.B., Tan, G.C.B. & Chen, V.H. (2006). Usability evaluation for mobile device: a comparison of laboratory and field tests. In: *MobileHCI '06: Proceedings of the 8th conference on Human-computer interaction with mobile devices and services.* Helsinki, Finland, New York, NY, USA: ACM. pp. 181-186.

Dunphy, P. & Yan, J. (2007). Do background images improve draw a secret graphical passwords? In: *CCS '07: Proceedings of the 14th ACM Conference on Computer and Communications Security.* Alexandria, Virginia, USA, New York, NY, USA: ACM. pp. 36-47.

Dybå, T., Kampenes, V.B. & Sjøberg, D.I.K. (2006). A systematic review of statistical power in software engineering experiments. *Information and Software Technology.* 48 (8): 745-755.

Eljetlawi, A.M. & Ithnin, N. (2008). Graphical Password: Comprehensive Study of the Usability Features of the Recognition Base Graphical Password Methods. In: *Third International Conference on Convergence and Hybrid Information Technology.* Busan, 11-13 November 2008, pp. 1137-1143.

Eljetlawi, A.M. (2010). Graphical password: Existing recognition base graphical password usability. In: *6th International Conference on Networked Computing.* Gyeongju, Korea (South), 14 June 2010, pp. 1-5.

Elmasri, R. & Navathe, S.B. (2011). *Fundamentals of Database Systems.* 6th edn. Pearson Education Inc.

Everitt, K.M., Bragin, T., Fogarty, J. & Kohno, T. (2009). A comprehensive study of frequency, interference, and training of multiple graphical passwords. In: *Proceedings of the 27th international conference on Human factors in computing systems.* CHI '09. Boston, MA, USA, New York, NY, USA: ACM. pp. 889-898.

Fiotakis, G., Raptis, D. & Avouris, N. (2009). Considering Cost in Usability Evaluation of Mobile Applications: Who, Where and When. In: *Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part I.* INTERACT '09. Uppsala, Sweden, Berlin, Heidelberg: Springer-Verlag. pp. 231-234.

Frøkjær, E., Hertzum, M. & Hornbæk, K. (2000). Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In: *CHI '00, The SIGCHI conference on Human factors in computing systems.* The Hague, The Netherlands, New York, NY, USA: ACM. pp. 345-352.

Fun, T.S., Beng, L.Y., Likoh, J. & Roslan, R. (2008). A lightweight and private mobile payment protocol by using mobile network operator. In: *International Conference on Computer and Communication Engineering, 2008.* Kuala Lumpur, 13-15 May 2008, pp. 162-166.

Furnell, S.M. (2007). Making security usable: Are things improving? *Computers & Security.* 26 (6): 434-443.

Gahtani, A.S.S. & King, M. (1999). Attitudes, satisfaction and usage: factors contributing to each in the acceptance of information technology. *Behaviour & Information Technology.* 18 (4): 277-297.

Goldberg, J., Hagman, J. & Sazawal, V. (2002). Doodling our way to better authentication. In: *CHI '02: Extended Abstracts on Human Factors in Computing Systems.* Minneapolis, Minnesota, USA, New York, NY, USA: ACM. pp. 868-869.

Gong, J. & Tarasewich, P. (2004). Guidelines for handheld mobile device interface design. In: *Proceedings of the 2004 DSI Annual Meeting.*

Google Inc (2011). *Unwrapping Ice Cream Sandwich on the Galaxy Nexus.* Available from: http://googleblog.blogspot.com/2011/10/unwrapping-ice-cream-sandwich-on-galaxy.html. [Accessed: 27/10/2011].

Gorlenko, L. & Merrick, R. (2003). No wires attached: Usability challenges in the connected mobile world. *IBM Systems Journal.* 42 (4): 639-651.

Green, A. (2007). Management of security policies for mobile devices. In: *4th Annual Conference on information Security Curriculum Development.* Kennesaw, Georgia, September 28th, New York: ACM.

Greene, S. & Finnegan, J. (2003). Usability of mobile devices and intelligently adapting to a user's needs. In: *ISICT '03: Proceedings of the 1st international symposium on Information and communication technologies.* Dublin, Ireland, Trinity College Dublin. pp. 175-180.

GrIDsure (2011). *GrIDsure - Pattern Based Authentication.* Available from: http://www.gridsure.com/. [Accessed: 12/08/2011].

Gritzalis, D. & Katsikas, S. (1996). Towards a formal system-to-system authentication protocol. *Computer Communications.* 19 (12): 954-961(8).

Gross, S. (2006). Towards Cooperative Self-Protecting Mobile Devices using Trustful Relationships. In: *SecureComm 06: International Conference on Security and Privacy in Communications Networks.* Baltimore, Maryland, USA, 28 Aug - 1 Sep, IEEE Computer Society. pp. 1-7.

Hager, C.T. & Midkiff, S.F. (2003). An analysis of Bluetooth security vulnerabilities. In: *IEEE Wireless Communications and Networking.* New Orleans, LA, USA, 20th March 2003, pp. 1825-1831.

Halpert, B.J. (2005). Authentication interface evaluation and design for mobile devices. In: *InfoSecCD '05: Proceedings of the 2nd annual conference on Information security curriculum development.* Kennesaw, Georgia, New York, NY, USA: ACM. pp. 112-117.

Hamilton, S.S., Carlisle, M.C. & Hamilton, J.A. (2007). A Global Look at Authentication. In: *IEEE SMC Information Assurance and Security Workshop, 2007.* West Point, NY, 20-22 June 2007, IEEE SMC. pp. 1-8.

Hashemi, M.R. & Soroush, E. (2006). A Secure m-Payment Protocol for Mobile Devices. In: *Canadian Conference on Electrical and Computer Engineering.* Ottawa, Canada, 7-10 May, IEEE Computer Society. pp. 294-297.

Hayashi, E., Dhamija, R., Christin, N. & Perrig, A. (2008). Use Your Illusion: Secure Authentication Usable Anywhere. In: *SOUPS '08, Proceedings of the 4th Symposium on Usable Privacy and Security.* Pittsburgh, Pennsylvania, New York, NY, USA: ACM. pp. 35-45.

Hazen, T.J., Weinstein, E., Heisele, B., Park, A. & Ming, J. (2006). Multi-Modal Face and Speaker Identification for Mobile Devices. : 123-138.

He, Q., Wu, D. & Khosla, P. (2004). Quest for Personal Control over Mobile Location Privacy. *IEEE Communications.* 42 : 130-136.

Heckle, R.R., Patrick, A. & Ozok, A. (2007). Perception and Acceptance of Fingerprint Biometric Technology. In: *SOUPS '07: Proceedings of the 3rd symposium on Usable privacy and security.* Pittsburgh, Pennsylvania, New York, NY, USA: ACM. pp. 154.

Heo, J., Ham, D., Park, S., Song, C. & Yoon, W.C. (2009). A framework for evaluating the usability of mobile phones based on multi-level, hierarchical model of usability factors. *Interacting with Computers.* 21 (4): 263-275.

Herzberg, A. & Margulies, R. (2012). Training Johnny to Authenticate (Safely). *IEEE Security & Privacy.* PP (1): 1-45.

Herzberg, A. & Margulies, R. (2011). *A Usable and Phishing-Resistant Fallback-Authentication Mechanism: Design and a User Study.* Bar Ilan University: Department of Computer Science.

HHS Web Communications Division (2009). *The Research-Based Web Design & Usability Guidelines.* Available from: http://www.usability.gov/guidelines/. [Accessed: 26/05/2011].

Hollingworth, H.L. (1913). Characteristic Differences between Recall and Recognition. *The American Journal of Psychology.* 24 (4): 532-544.

Hong, D., Man, S., Hawes, B. & Matthews, M. (2004). A Graphical Password Scheme Strongly Resistant to Spyware. In: *Proceedings of the International Conference on Security and Management.* Las Vegas, NV, 21-24 June, CSREA Press. pp. 94-100.

Hossain, A., Jahan, S., Hussain, M.M., Amin, M.R. & Shah Newaz, S.H. (2008). A proposal for enhancing the security system of short message service in GSM. In: *2nd International Conference on Anti-counterfeiting, Security and Identification.* Guiyang, China, 20-23 Aug, IEEE Computer Society. pp. 235-240.

Programming stuff. (2007). [Online]. Available from: http://www.the-interweb.com/serendipity/index.php?/archives/94-A-brief-analysis-of-40,000-leaked-MySpace-passwords.html. [Accessed: 10/11/2011].

Hubaux, J., Buttyán, L. & Capkun, S. (2001). The quest for security in mobile ad hoc networks. In: *MobiHoc '01: 2nd ACM International Symposium on Mobile Ad Hoc Networking and Computing.* Long Beach, CA, USA, 4-5 October, New York, NY, USA: ACM. pp. 146-155.

Huifang, H., Xinshen, J. & Guangqiang, L. (2008). A Novel Access Authentication Scheme Based on ECC for 3G-WLAN Interworking Network. In: *International Conference on Computer Science and Software Engineering.* Wuhan, China, 12-14 December 2008, IEEE Computer Society. pp. 1237-1241.

Humm, A., Hennebert, J. & Ingold, R. (2009). Combined Handwriting and Speech Modalities for User Authentication. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans.* 39 (1): 25-35.

Ikram, M., Chowdhury, A.H., Redwan, H., Koh, J., Kim, K. & Kim, D. (2008). A Lightweight Mutual Authentication Scheme for Mobile Radio Frequency Identification (mRFID) Systems. In: *IEEE International Performance, Computing anf Communications Conference.* Austin, Texas, 7-10 July 2008, IEEE Computer Society. pp. 289-296.

Inglesant, P.G. & Sasse, M.A. (2010). The true cost of unusable password policies: password use in the wild. In: *Proceedings of the 28th international conference on Human factors in computing systems.* CHI '10. Atlanta, Georgia, USA, New York, NY, USA: ACM. pp. 383-392.

Ion, I., Dragovic, B. & Crispo, B. (2007). Extending the Java Virtual Machine to Enforce Fine-Grained Security Policies in Mobile Devices. In: *23rd Annual Computer Security Applications Conference.* Miami Beach, Florida, USA, 10-14 December, IEEE Computer Society. pp. 233-242.

ISO/IEC (1999). *Human-centred design processes for interactive systems.* 13407. ISO/IEC.

ISO/IEC (1998). *9241-11 Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on Usability.* ISO/IEC.

Jambon, F., Golanski, C. & Pommier, P. (2007). Meta-Evaluation of a Context-Aware Mobile Device Usability. In: *UBICOMM '07: Proceedings of the International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies.* Washington, DC, USA: IEEE Computer Society. pp. 21-26.

Jamil, T. (2004). The Rijndael algorithm. *Potentials, IEEE.* 23 (2): 36-38.

Jeong, J., Shin, D. & Shin, D. (2004). An XML-based single sign-on scheme supporting mobile and home network service environments. *IEEE Transactions on Consumer Electronics.* 50 (4): 1081-1086.

Jeong, Y., Yun, Y., Jung, B. & Kim, K. (2008). Protection Profile for Security Enhancement of Embedded Operating System for Mobile Terminals. In: *10th International Conference on Advanced Communication Technology.* pp. 1908-1911.

Jermyn, I., Mayer, A., Monrose, F., Reiter, M.K. & Rubin, A.D. (1999). The design and analysis of graphical passwords. In: *SSYM'99: Proceedings of the 8th conference on USENIX Security Symposium.* Washington, D.C., Berkeley, CA, USA: USENIX Association. pp. 1-1.

Jeyaraman, S. & Topkara, U. (2005). Have the cake and eat it too - infusing usability into text-password based authentication systems. In: *21st Annual Computer Security Applications Conference.* Tucson, AZ, 05-09 December 2005, pp. 10 pp.-482.

Johnston, J., Eloff, J.H.P. & Labuschagne, L. (2003). Security and human computer interfaces. *Computers & Security.* 22 (8): 675-684.

Jokela, T., Iivari, N., Matero, J. & Karukka, M. (2003). The Standard of User-Centered Design and the Standard Definition of Usability: Analyzing ISO 13407 against ISO 9241-11. In: *Proceedings of the Latin American conference on Human-Computer Interaction.* CLIHC '03. Rio de Janeiro, Brazil, New York, NY, USA: ACM. pp. 53-60.

Kaikkonen, A. & Roto, V. (2003). Navigating in a mobile XHTML application. In: *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems.* Ft. Lauderdale, Florida, USA, New York, NY, USA: ACM. pp. 329-336.

Kainda, R., Flechais, I. & Roscoe, A.W. (2009). Usability and security of out-of-band channels in secure device pairing protocols. In: *SOUPS '09: Proceedings of the 5th Symposium on Usable Privacy and Security.* Mountain View, California, New York, NY, USA: ACM. pp. 1-12.

Kapadia, A. (2007). A Case (Study) For Usability in Secure Email Communication. *IEEE Security and Privacy.* 5 (2): 80-84.

Karlog, C., Sastry, N. & Wagner, D. (2004). Tinysec: A link layer security architecture for wireless. In: *ACM SenSys 2004.* Baltimore, Maryland, US, November 3-5, ACM. pp. 162-174.

Kelly, M. (2012). *eHarmony Password Dump Analysis.* [Online] 25/06/2012. Available from: http://blog.spiderlabs.com/2012/06/eharmony-password-dump-analysis.html. [Accessed: 26/06/2012].

Killourhy, K. & Maxion, R. (2008). The Effect of Clock Resolution on Keystroke Dynamics. In: Lippmann, R. et al. (ed.) *Recent Advances in Intrusion Detection.* Lecture Notes in Computer Science. vol. 5230. Springer Berlin / Heidelberg.

Kim, J., Choi, D., Kim, I. & Kim, H. (2006). Product Authentication Service of Consumer's mobile RFID Device. In: *ISCE '06, IEEE Tenth International Symposium on Consumer Electronics.* St. Petersburg, pp. 1-6.

Kim, J. & Kim, H. (2006). Security Vulnerability and Considerations in Mobile RFID environment. In: *ICACT 2006, 8th International Conference Advanced Communication Technology.* pp. 801-804.

Kim, Y.B. (2010). Real-time Analysis of Time-based Usability and Accessibility for Human Mobile-Web Interactions in the Ubiquitous Internet. *Journal of Universal Computer Science.* 16 (15): 1953-1972.

Kitchenham, B.A. & Charters, S.M. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering, Version 2.3.* Technical report. Keele University and Durham University: .

Kitchenham, B., Brereton, P. & Budgen, D. (2012). Mapping study completeness and reliability - a case study. In: *Evaluation & Assessment in Software Engineering (EASE 2012), 16th International Conference on.* pp. 126-135.

Kjeldskov, J. & Stage, J. (2004). New techniques for usability evaluation of mobile systems. *International Journal of Human-Computer Studies.* 60 (5-6): 599-620.

Klockar, T., Carr, D.A., Hedman, A., Johansson, T. & Bengtsson, F. (2003). Usability of mobile phones. In: *In Proceeding of the 19th International Symposium on Human Factors in Telecommunications.* Luleå University of Technology, SE-971 87 Luleå, Sweden: pp. 197-204.

Koblitz, N. (1987). Elliptic Curve Cryptosystems. *Mathematics of Computation.* 48 (177): 203-209.

Kolsi, O. & Virtanen, T. (2004). MIDP 2.0 security enhancements. In: *Proceedings of the 37th Annual Hawaii International Conference on System Sciences.* Big Island, Hawaii, USA, 05-08 January, IEEE Computer Society. pp. 8 pp.

Kong, J., Hong, X., Yi, Y., Park, J., Liu, J. & Gerla, M. (2005). A secure ad-hoc routing approach using localized self-healing communities. In: *MobiHoc '05, Proceedings of the 6th ACM international symposium on Mobile ad hoc networking and computing.* Urbana-Champaign, IL, USA, 25-28 May, New York, NY, USA: ACM. pp. 254-265.

Kowalski, S. & Goldstein, M. (2006). Consumers' Awareness of, Attitudes Towards and Adoption of Mobile Phone Security. In: *Human Factors in Telecommunication (HFT 06).* 20-23 March, .

Kuntze, N. & Schmidt, A.U. (2006). Trusted computing in mobile action. In: *Proceedings of the Information Security South Africa (ISSA) 2006 From Insight to Foresight Conference.* Sandton, South Africa, 05-07 July, .

Lashkari, A.H., Towhidi, F., Saleh, R. & Farmand, S. (2009). A Complete Comparison on Pure and Cued Recall-Based Graphical User Authentication Algorithms. In: *Second*

International Conference on Computer and Electrical Engineering. ICCEE '09. Dubai, 28-30 December, pp. 527-532.

Lee, S., Suh, S., Jeong, B. & Mo, S. (2008). A Multi-Layer Mandatory Access Control Mechanism for Mobile Devices Based on Virtualization. In: *5th IEEE Consumer Communications and Networking Conference.* Las Vegas, Nevada, USA, 10-12 January, IEEE Computer Society. pp. 251-256.

Leung, A. & Mitchell, C. (2007). Ninja: Non Identity Based, Privacy Preserving Authentication for Ubiquitous Environments. In: *Proceedings of the 9th International Conference on Ubiquitous Computing.* UbiComp '07. Innsbruck, Austria, pp. 73-90.

Lewis, J.R. (1995). IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction.* 7 (1): 57-78.

Liang, W. & Wang, W. (2004). A local authentication control scheme based on AAA architecture in wireless networks. In: *IEEE 60th Vehicular Technology Conference.* Los Angeles, California, USA, 26-29 September, IEEE Computer Society. pp. 5276-5280.

Lim, Y., Pangam, A., Periyasami, S. & Aneja, S. (2006). Comparative analysis of high- and low-fidelity prototypes for more valid usability evaluations of mobile devices. In: *Proceedings of the 4th Nordic conference on Human-computer Interaction.* NordiCHI '06. Oslo, Norway, New York, NY, USA: ACM. pp. 291-300.

Lin, X., Lu, R., Ho, P., Shen, X. & Cao, Z. (2007). A Novel Compromise-Resilient Authentication System for Wireless Mesh Networks. In: *IEEE Wireless Communications and Networking Conference 2007.* Hong Kong, China, 11-15 March 2007, IEEE Computer Society. pp. 3541-3546.

Linlin Shen, Nengheng Zheng, Songhao Zheng & Wei Li (2010). Secure mobile services by face and speech based personal authentication. In: *IEEE International Conference on Intelligent Computing and Intelligent Systems.* ICIS. Xiamen, 29-31 October, pp. 97-100.

Lockhart, R.S., Craik, F.I. & Jacoby, L. (1976). Depth of processing, recognition and recall. In: Anonymous *Recall and recognition.* Oxford, England: John Wiley & Sons.

Lowe, G. (1997). A hierarchy of authentication specifications. In: *Proceedings of the 10th Computer Security Foundations Workshop.* Rockport, MA, USA, 10-12 Jun 1997, pp. 31-43.

Mack, Z. & Sharples, S. (2009). The importance of usability in product choice: A mobile phone case study. *Ergonomics.* 52 (12): 1514-1528.

Madlmayr, G. (2008). A mobile trusted computing architecture for a near field communication ecosystem. In: *iiWAS '08, 10th International Conference on Information Integration and Web-based Applications & Services.* Linz, Austria, New York, NY, USA: ACM. pp. 563-566.

Madlmayr, G., Dillinger, O., Langer, J. & Scharinger, J. (2008). Management of Multiple Cards in NFC-Devices. In: *Eighth Smart Card Research and Advanced Applications Conference.* London, UK, 8-11 September, Berlin, Heidelberg: Springer-Verlag. pp. 149-161.

Man, S., Hong, D. & Matthews, M. (2003). A shoulder-surfing resistant graphical password scheme. In: *International conference on security and management.* Las Vegas, June, pp. 105-111.

Markova, M. & Aula, A. (2007). Conceptualizing How Usability of Mobile Services Affects Business Performance. In: *International Conference on the Management of Mobile Business, 2007.* Toronto, Ont., USA, 9-11 July, pp. 36-43.

Masrom, M., Towhidi, F. & Lashkari, A.H. (2009). Pure and cued recall-based graphical user authentication. In: *International Conference on Application of Information and Communication Technologies.* Baku, 14-16 October, pp. 1-6.

Mayrhofer, R. (2006). A context authentication proxy for IPSec using spatial reference. In: *TwUC 2006, 1st International Workshop on Trustworthy Ubiquitous Computing.* pp. 449-462.

Mayrhofer, R. & Gellersen, H. (2007). Shake well before use: Authentication based on accelerometer data. In: *International Conference on Pervasive Computing.* Toronto, Ontario, Canada, 13-26 May, Springer. pp. 144-161.

Me, G., Pirro, D. & Sarrecchia, R. (2006). A mobile based approach to strong authentication on Web. In: *ICCGI '06: International Multi-Conference on Computing in the Global Information Technology.* Bucharest, Romania, 01 August, IEEE Computer Society. pp. 67-67.

Mihajlov, M., Jerman-Blazic, B. & Ilievski, M. (2011). ImagePass - Designing graphical authentication for security. In: *7th International Conference on Next Generation Web Services Practices (NWeSP).* Salamanca, 19-21 October, pp. 262-267.

Mihajlov, M., Jerman-Blazic, B. & Ilievski, M. (2011). Recognition-Based Graphical Authentication with Single-Object Images. In: *Developments in E-systems Engineering (DeSE), 2011.* Dubai, 6-8 December, pp. 203-208.

Mihajlov, M., Jerman-Blazic, B. & Josimovski, S. (2011). A conceptual framework for evaluating usable security in authentication mechanisms - usability perspectives. In: *5th International Conference on Network and System Security (NSS).* Milan, 6-8 September, pp. 332-336.

Mika, P.K. & Röykkee, M. (2001). The Three Facets of Usability In Mobile Handsets. In: *Proceeding of CHI 2001, Workshop, Mobile Communications: Understanding Users, Adoption & Design Sunday and Monday.* ACM.

Miller, G.A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review.* 101 (2): 343-352.

Mobbeel *BioWallet Signature - Biometric Mobile Security.* Available from: http://www.mobbeel.com/products/biowallet. [Accessed: 12/01/2012].

Moncur, W. & Leplâtre, G. (2007). Pictures at the ATM: exploring the usability of multiple graphical passwords. In: *Proceedings of the SIGCHI conference on Human factors in computing systems.* CHI '07. San Jose, California, USA, New York, NY, USA: ACM. pp. 887-894.

Moody, J. (2004). Public perceptions of biometric devices: The effect of misinformation on acceptance and use. *Journal of Issues in Informing Science and Information Technology.* : 753-761.

Mufti, M. & Khanum, A. (2004). Design and implementation of a secure Mobile IP protocol. In: *International Networking and Communication Conference.* INCC 2004. 11-13 June, pp. 53-57.

Nali, D. & Thorpe, J. (2004). *Analyzing User Choice in Graphical Passwords.* University of Ottawa: School of Information Technology and Engineering.

Narayanaswamy, S., Jianying Hu & Kashi, R. (1999). User interface for a PCS smart phone. In: *IEEE International Conference on Multimedia Computing and Systems.* Florence, 07-11 June 1999, IEEE. pp. 777-781.

Nguessan, D. & Martini, J.S.C. (2008). Study and Implementation of a Solution to Security Management for Mobile Environments Based on Tuple. In: *The 9th International Conference for Young Computer Scientists.* ICYCS 2008. Hunan, 18-21 November 2008, pp. 2014-2020.

Nielsen, J. (2005). *Ten Usability Heuristics.* Available from: http://www.useit.com/papers/heuristic/heuristic_list.html.

Nielsen, J. & Levy, J. (1994). Measuring usability: preference vs. performance. *Communications of the ACM.* 37 (4): 66-75.

Noponen, S. & Karppinen, K. (2008). Information Security of Remote File Transfers with Mobile Devices. In: *32nd Annual IEEE International Computer Software and Applications,* COMPSAC '08. Turku, Finland, 28 July - 1 August, IEEE Computer Society. pp. 973-978.

Norman, D.A. (1988). *The Design of Everyday Things.* Basic Books.

Oppermann, R. (2002). User-interface Design. In: Adelsberger, H. et al. (ed.) *Handbook on information technologies for education and training.* Berlin: Springer.

Ou, C., Ou, C.R. & Wang, Y. (2008). Security of Mobile Agent-Based Web Applications. In: *APSCC '08: IEEE Asia-Pacific Services Computing Conference.* Yilan, Taiwan, 9-12 December, IEEE Computer Society. pp. 107-112.

Palomar, E., Tapiador, J.M.E., Hernandez-Castro, J.C. & Ribagorda, A. (2007). Dealing with Sporadic Strangers, or the (Un)Suitability of Trust for Mobile P2P Security. In: *18th*

International Workshop on Database and Expert Systems Applications. DEXA '07. Regensburg, Germany, 3-7 September, Washington, DC, USA: IEEE Computer Society. pp. 779-783.

Pan, T., Zheng, L., Fang, C., Huang, W. & Fang, L. (2008). A New Mobile Information Security Solution Based on External Electronic Key. In: *ICMeCG '08, International Conference on Management of e-Commerce and e-Government.* Nanchang, China, 17-19 October, IEEE Computer Society. pp. 25-29.

Park, K., Choi, H.J. & Park, K.H. (2005). An interoperable authentication system using zigbee-enabled tiny portable device and PKI. In: *International Conference on Next Generation PC.*

Park, N., Lee, J., Kim, H., Chung, K. & Sohn, S. (2006). A Layered Approach to Design of Light-Weight Middleware Systems for Mobile RFID Security (SMRM : Secure Mobile RFID Middleware System). In: *10th IEEE/IFIP Network Operations and Management Symposium.* NOMS 2006. Vancouver, BC, USA, 3-7 April 2006, IEEE. pp. 1-4.

Payne, B.D. & Edwards, W.K. (2008). A Brief Introduction to Usable Security. *IEEE Internet Computing.* 12 (3): 13-21.

Perelson, S. & Botha, R.A. (2004). An Investigation into Access Control for Mobile Devices. In: *Information Security South Africa Enabling Tomorrow Conference.* Gallagher Estate, Midrand, South Africa, 30 June – 2 July, ISSA.

Podio, F.L. (2002). Personal authentication through biometric technologies. In: *IEEE 4th International Workshop on Networked Appliances.* Gaithersburg, 15-16 January 2002, pp. 57-66.

Priya, B.S. & Rajesh, R. (2011). A note on fingerprint recognition systems. In: *3rd International Conference on Electronics Computer Technology*. Kanyakumari, 8-10 April 2011, pp. 95-98.

Raaijmakers, J.G. & Shiffrin, R.M. (1992). Models for recall and recognition. *Annual Review of Psychology.* 43 : 205-234.

Ravishankar, B. & Harishankar, M.V. (2008). Roaming Issues in 3GPP Security Architecture and Solution Using UMM Architecture. In: *UBICOMM '08: Second International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies.* Valencia, Spain, 29 September - 4 October, IEEE Computer Society. pp. 457-462.

Real User Corporation (2004). *The Science Behind Passfaces.*

Real User Corporation (2001). *PKI and Passfaces™: Synergistic or competitive.* 2010.

Renaud, K.V. (2009). Guidelines for designing graphical authentication mechanism interfaces. *International Journal of Information and Computer Security.* 3 (1): 60-85.

Renaud, K.V. & Smith, E. (2001). Helping Users to Remember Their Passwords. In: *Annual Conference of the South African Institute of Computer Scientists and Information Technologies.* Pretoria, South Africa, .

Rock, I. & Engelstein, P. (1959). A Study of Memory for Visual Form. *The American Journal of Psychology.* 72 (2): 221-229.

Rosnow, R.L. & Rosenthal, R. (1997). *People Studying People: Artifacts and Ethics in Behavioral Research.* 1st edn. USA: Freeman.

Sasamoto, H., Christin, N. & Hayashi, E. (2008). Undercover: authentication usable in front of prying eyes. In: *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems.* CHI '08. Florence, Italy, New York, NY, USA: ACM. pp. 183-192.

Sastry, N., Shankar, U. & Wagner, D. (2003). Secure verification of location claims. In: *WiSe '03: 2nd ACM workshop on Wireless security.* San Diego, California, USA, 19 September, New York, NY, USA: ACM. pp. 1-10.

Savola, R. & Holappa, J. (2005). Self-measurement of the information security level in a monitoring system based on mobile ad hoc networks. In: *IEEE International Workshop on Measurement Systems for Homeland Security, Contraband Detection and Personal Safety Workshop.* pp. 42-49.

Saxena, N., Uddin, M.B., Voris, J. & Asokan, N. (2011). Vibrate-to-unlock: Mobile phone assisted user authentication to multiple personal RFID tags. In: *IEEE International Conference on Pervasive Computing and Communications.* PerCom 2011. Seattle, WA, USA, 21-25 March, IEEE. pp. 181-188.

Schusteritsch, R., Wei, C.Y. & LaRosa, M. (2007). Towards the perfect infrastructure for usability testing on mobile devices. In: *Extended abstracts on Human factors in computing systems.* CHI '07. San Jose, CA, USA, New York, NY, USA: ACM. pp. 1839-1844.

SebastianApps *AppSeal.* Available from: http://www.sebastianapps.com/.

Senn, S. (2002). *Cross-over Trials in Clinical Research.* 2nd edn. Chichester, UK: John Wiley & Sons Ltd.

Shepard, R.N. (1967). Recognition Memory for Words Sentences and Pictures. *Journal Of Verbal Learning And Verbal Behavior.* 6 (1): 156-163.

Shinder, D. (2005). *Bluetooth: Is it a Security Threat?* [Online] 09/08/2005. Available from: http://www.windowsecurity.com/articles-tutorials/Wireless_Security/Bluetooth-Security-Threat.html. [Accessed: 29/07/2013].

Siek, K.A., Rogers, Y. & Connelly, K.H. (2005). Fat Finger Worries: How Older and Younger Users Physically Interact with PDAs. *Lecture notes in Computer Science.* 2585/2005 (267): 280.

Sirlantzis, K., Howells, G., Deravi, F., Hoque, S., Radu, P., McConnon, G., Savatier, X., Ertaud, J.-., Ragot, N., Dupuis, Y. & Iraqui, A. (2010). Nomad Biometric Authentication: Towards Mobile and Ubiquitous Person Identification. In: *International Conference on Emerging Security Technologies.* EST 2010. Canterbury, 6-7 September, pp. 1-6.

Smetters, D.K. & Grinter, R.E. (2002). Moving from the design of usable security technologies to the design of useful secure applications. In: *Proceedings of the 2002 workshop on New security paradigms.* NSPW '02. Virginia Beach, Virginia, New York, NY, USA: ACM. pp. 82-89.

Smith, S.W. (2003). Humans in the loop: human-computer interaction and security. *IEEE Security and Privacy.* 1 (3): 75-79.

Sobrado, L. & Birget, J. (2002). *Graphical Passwords.* The Rutgers Scholar, An Electronic Bulletin for Undergraduate Research: Rutgers.

Spafford, E.H. (1992). Observing Reusable Password Choices. In: *Proceedings of the 3rd USENIX Security Symposium.* September 1992, .

Standing, L. (1973). Learning 10,000 pictures. *The Quarterly Journal of Experimental Psychology.* 25 (2): 207-222.

Stubblefield, A. & Simon, D. (2004). *Inkblot Authentication.* MSR-TR-2004-85. Microsoft Research.

Sun, D., Huai, J., Sun, J., Zhang, J. & Feng, Z. (2008). A new design of wearable token system for mobile device security. *A new design of wearable token system for mobile device security.* 54 (4): 1784-1789.

Sun, K., Xu, R., Deng, J., Haynes, L., Li, J.H., Gruenwald, L., Sanchez, C., Weber, G. & Mayhew, M.J. (2008). Securing MANET databases using metadata and context information. In: *MILCOM 2008, IEEE Military Communications Conference.* pp. 1-6.

Suo, X., Zhu, Y. & Owen, G.S. (2005). Graphical Passwords: A Survey. In: *21st Annual Computer Security Applications Conference.* pp. 10 pp.-472.

Symantec Corporation (2012). *The Symantec Smartphone Honey Stick Project.* Available from: http://www.symantec.com/content/en/us/about/presskits/b-symantec-smartphone-honey-stick-project.en-us.pdf.

Tafasa (2011). *PatternLock.* Available from: http://www.tafasa.com/patternlock.html. [Accessed: 12/08/2011].

Takada, T. & Koike, H. (2003). Awase-E: Image-Based Authentication for Mobile Phones Using User's Favorite Images. In: *5th International Symposium on Human-Computer Interaction with Mobile Devices and Services.* Berlin: Springer Verlag. pp. 347-351.

Tari, F., Ozok, A.A. & Holden, S.H. (2006). A comparison of perceived and real shoulder-surfing risks between alphanumeric and graphical passwords. In: *Proceedings of the*

second symposium on Usable privacy and security. SOUPS '06. Pittsburgh, Pennsylvania, New York, NY, USA: ACM. pp. 56-66.

TAT (2011). *Augmented ID.* Available from: http://www.tat.se/videos/. [Accessed: 15/08/2011].

Tichy, W.F., Lukowicz, P., Prechelt, L. & Heinz, E.A. (1995). Experimental Evaluation in Computer-Science - a Quantitative Study. *Journal of Systems and Software*. 28 (1): 9-18.

Toa, H. & Adams, C. (2008). Pass-Go: A Proposal to Improve the Usability of Graphical Passwords. *International Journal of Network Security*. 7 (2): 273-292.

Towhidi, F. & Masrom, M. (2009). A Survey on Recognition Based Graphical User Authentication Algorithms. *International Journal of Computer Science and Information Security*. 6 (2): 119-127.

Trusted Computing Group (2005). *TCG Architecture Overview, Version 1.4.* Available from: http://www.trustedcomputinggroup.org/resources/tcg_architecture_overview_version_14.

Tulving, E. & Watkins, M.J. (1973). Continuity between Recall and Recognition. *The American Journal of Psychology*. 86 (4): 739-748.

van Thanh, D., Jorstad, I., Engelstad, P., Jonvik, T. & Feng, B. (2008). Authentication in a Multi-access IMS Environment. In: *IEEE International Conference on Wireless and Mobile Computing, Networking and Communications.* Crete Island, Greece, 6-8 August, IEEE Computer Society. pp. 613-618.

Wang, G., Wang, Q., Cao, J. & Guo, M. (2007). An Effective Trust Establishment Scheme for Authentication in Mobile Ad Hoc Networks. In: *7th IEEE International Conference on Computer and Information Technology*. CIT 2007. Fukushima, Japan, 16-19 October, IEEE Computer Society. pp. 749-754.

Wang, H., Cao, J. & Zhang, Y. (2002). Ticket-based service access scheme for mobile users. In: *ACSC '02, Proceedings of the twenty-fifth Australasian conference on Computer science.* Melbourne, Australia, January/February, Darlinghurst, Australia, Australia: Australian Computer Society, Inc. pp. 285-292.

Wangensteen, A., Lunde, L., Jørstad, I. & van Thanh, D. (2006). A Generic Authentication System based on SIM. In: *ICISP '06: International Conference on Internet Surveillance and Protection.* Côte d'Azur, France, 27-29 August, Cap Esterel, Côte d'Azur, France: IEEE Computer Society. pp. 24-24.

Weinshall, D. (2006). Cognitive Authentication Schemes Safe Against Spyware. In: *IEEE Symposium on Security and Privacy*. Berkeley/Oakland, CA, 21-24 May, IEEE. pp. 295-300.

Weir, C.S., Douglas, G., Richardson, T. & Jack, M. (2010). Usable security: User preferences for authentication methods in eBanking and the effects of experience. *Interacting with Computers.* 22 (3): 153-164.

Weiss, R. & De Luca, A. (2008). PassShapes: Utilizing Stroke Based Authentication to Increase Password Memorability. In: *Proceedings of the 5th Nordic conference on Human-computer interaction.* NordiCHI '08. Lund, Sweden, New York, NY, USA: ACM. pp. 383-392.

Weiss, S. (2002). *Handheld Usability.* New York, NY, USA: John Wiley & Sons, Inc.

Whitten, A. & Tygar, J.D. (1999). Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0. In: *8th USENIX Security Symposium.*

Wiedenbeck, S., Waters, J., Birget, J., Brodskiy, A. & Memon, N. (2005). PassPoints: design and longitudinal evaluation of a graphical password system. *International Journal of Human-Computer Studies.* 63 (1-2): 102-127.

Wood, H.M. (1977). *Use of Passwords for Controlled Access to Computer Resources.* Washington, D.C.: Superintendent of Documents, U.S. Government Printing Office.

Wright, T., Yoong, P., Noble, J., Cliffe, R., Hoda, R., Gordon, D. & Andreae, C. (2005). Usability methods and mobile devices: an evaluation of MoFax. In: *MUM '05: Proceedings of the 4th international conference on Mobile and ubiquitous multimedia.* Christchurch, New Zealand, New York, NY, USA: ACM. pp. 26-33.

Yan, J., Blackwell, A., Anderson, R. & Grant, A. (2004). Password memorability and security: empirical results. *IEEE Security and Privacy.* 2 (5): 25-31.

Yan, J. & El Ahmad, A.S. (2008). Usability of CAPTCHAs or usability issues in CAPTCHA design. In: *Proceedings of the 4th Symposium on Usable Privacy and Security.* SOUPS '08. Pittsburgh, Pennsylvania, New York, NY, USA: ACM. pp. 44-52.

Yee, K. (2004). Aligning Security and Usability. *IEEE Security and Privacy.* 2 (5): 48-55.

Yee, K. (2002). User Interaction Design for Secure Systems. In: *ICICS '02: Proceedings of the 4th International Conference on Information and Communications Security.* London, UK: Springer-Verlag. pp. 278-290.

Yu, S., Zhang, Y., Song, C. & Chen, K. (2004). A security architecture for Mobile Ad Hoc Networks. In: *18th APAN Meetings.*

Yuan Fu Qiu, Yoon Ping Chui & Helander, M.G. (2006). Usability Analysis of Mobile Phone Camera Software Systems. In: *IEEE Conference on Cybernetics and Intelligent Systems.* Bangkok, 7-9 June, IEEE. pp. 1-6.

Yung-Wei Kao, Hui-Zhen Gu & Shyan-Ming Yuan (2008). Personal Based Authentication by Face Recognition. In: *Fourth International Conference on Networked Computing and Advanced Information Management.* NCM '08. Gyeongju, 2-4 September, pp. 581-585.

Zhao, H. & Li, X. (2007). S3PAS: A Scalable Shoulder-Surfing Resistant Textual-Graphical Password Authentication Scheme. In: *International Conference on Advanced Information Networking and Applications Workshops.* AINAW '07. Niagara Falls, Ontario, USA, 21-23 May, Los Alamitos, CA, USA: IEEE Computer Society. pp. 467-472.

Zheng, Y., He, D., Tang, X. & Wang, H. (2005). AKA and Authorization Scheme for 4G Mobile Networks Based on Trusted Mobile Platform. In: *Fifth International Conference on Information, Communications and Signal Processing.* Bangkok, Thailand, 6-9 December, IEEE Computer Society. pp. 976-980.

Zheng, Y., Xia, J. & He, D. (2008). Trusted user authentication scheme combining password with fingerprint for mobile devices. In: *ISBAST 2008: International Symposium on Biometrics and Security Technologies.* Islamabad, Pakistan, 23-24 April, IEEE Computer Society. pp. 1-8.

Zviran, M. & Haga, W.J. (1999). Password security: an empirical study. *Journal of Management Information Systems.* 15 (4): 161-185.

Appendix A: Experiment Questionnaire

Project Title: An experiment into the usability of authentication models on mobile devices

Section 1: Personal Details

Gender: ❑ Male ❑ Female

Age:
❑ 18 to 21
❑ 22 to 34
❑ 35 to 44
❑ 45 to 54
❑ 55 to 64
❑ 65+
❑ Prefer not to say

Do you have any previous experience of using a Smartphone?
❑ Yes ❑ No

If yes, give a brief description

_____

_____

Do you own a mobile of any kind?
❑ Yes ❑ No

If yes, do you use a pin or other security feature (such as a graphical unlock pattern) on it to secure the device against other people using it?

_____

Order performed: ADP APD DAP DPA PAD PDA

Participant Code:

Section 2: Application details

1. Whenever I made a mistake using the system, I could recover quickly and easily.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

2. The instructions provided for the system were easy to understand.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

3. The organisation of the information on the system screens was clear.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

4. The use of colours in the application was distracting.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

5. I was able to create an account and log in quickly using this application.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

6. The application was responsive to my use of buttons or options.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

7. It was easy to learn how to use the application.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS

8. The small screen made the application hard to use.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

9. It was difficult to get through the login process.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

10. I liked using the interface of this application.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

11. Overall, it was not easy to use the application.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

12. This application has all the functions and capabilities I would expect it to have.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

13. Overall I am satisfied with this application.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

Appendix B: Ethical approval confirmation letters