DO HOUSES HAVE FACES?
THE EFFECT OF IMAGE TYPE IN RECOGNITION-BASED GRAPHICAL PASSWORDS

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of the requirement for the degree

Masters of Arts

by

Max Hlywa

Department of Psychology
Carleton University

December 2010

Your file Votre référence
ISBN: 978-0-494-79563-7
Our file Notre référence
ISBN: 978-0-494-79563-7

Abstract

Graphical passwords are a novel method of knowledge-based authentication that show promise for improved usability and memorability. This study examined the effect of image type in cognometric, recognition-based, graphical passwords. Specifically, the usability of such authentication schemes was explored at security levels equivalent to those acceptable for text passwords. Related theory was drawn upon to consider the relative strength of visual memory, to distinguish recognition from recall, and to explore face recognition by humans. The study consisted of three conditions, each using a different type of image (houses, faces, and objects). Image type was the independent variable, and login errors, login success, and login time were the dependent variables. The results showed that participants in the object images condition performed slightly better than those in the face images condition, and that participants in the house images condition had great difficulty. Importantly, there was little evidence of superior performance resulting from the use of face images in the authentication scheme.

Acknowledgements

To my supervisor Dr. Robert Biddle, whose caring and thoughtful nature continues to serve and inspire his students, and without whom this thesis would not have been possible. To my co-supervisor Dr. Andrew Patrick, whose insight and guidance was vital to the success of the thesis. To Chris Deschamps, who developed the authentication system used in the study. To my committee members, Dr. Warren Thorngate, Dr. Evelyn Maeder, Dr. Anil Somayaji, and Dr. Michael Wohl for their questions and criticism. To my friendly and helpful colleagues in both the Hotsoft and HOT labs. Last but not least, to my loving and patient family.

Table of Contents

Introduction 1
Graphical Passwords 3
Related Theory 7
Visual Memory 7
Recognition vs. Recall 8
Face Recognition 9
Password Space 13
Research Question 17
Research Study 17
Research Hypotheses 17
Method 20
Participants 20
Materials 20
Equipment 21
Procedure 22
First Session 23
Online Access 24
Second Session 24
Analysis Plan 24
Hypothesis One 25
Hypothesis Two 25
Hypothesis Three 25
Other Questions 26
Results 26
Hypothesis One 27
Hypothesis Two 32
Hypothesis Three 44
Questionnaire Feedback 55
Discussion 59
Conclusion 72
References 74

List of Tables

Table 1: Text password criteria and resulting theoretical space. 15
Table 2: Cognometric password criteria and resulting theoretical space 16
Table 3: Descriptives for login errors across conditions in each time period. 29
Table 4: Descriptives for login success across conditions in each time period 34
Table 5: Mean password memory time 40
Table 6: Descriptives for login time across conditions in each period. 45
Table 7: Transformed average login times, multiple comparisons 50
Table 8: Transformed average login times, multiple comparisons 55
Table 9: Mean ratings across conditions for the first questionnaire theme 56
Table 10: Mean ratings across conditions for the second questionnaire theme 57
Table 11: Mean ratings across conditions for the third questionnaire theme 57
Table 12: Mean ratings across conditions for the fourth questionnaire theme 58
Table 13: Mean ratings across conditions for the fifth questionnaire theme 58
Table 14: Mean ratings across conditions for the sixth questionnaire theme 59

List of Figures

Figure 1: Sample of the PassFaces login screen 3
Figure 2: Sample of the Draw-a-Secret (DAS) login screen, showing a user drawing. 4
Figure 3: A sample PassPoints image, showing user-selected click points. 5
Figure 4: Examples of the image types used in the study 18
Figure 5: Distribution of mean login errors during the 2nd lab session 30
Figure 6: Mean login errors during the 2nd lab session, separated by condition. 31
Figure 7: Distribution of login success scores during the 1st lab session. 35
Figure 8: Distribution of login success scores 1-2 days after the 1st lab session. 36
Figure 9: Distribution of login success scores 3-4 days after the 1st lab session. 37
Figure 10: Distribution of login success scores during the 2nd lab session 38
Figure 11: Distribution of mean memory time, including all conditions. 41
Figure 12: Distribution of mean memory time across conditions 42
Figure 13: Distribution of average login times during the 2nd lab session 46
Figure 14: Distributions of average login times during the 2nd lab session. 47
Figure 15: Distribution of transformed average login times during the 2nd lab session 48
Figure 16: Distributions of transformed average login times during the 2nd lab session 49
Figure 17: Distribution of average login time, all conditions & periods 51
Figure 18: Distributions of average login times, all conditions & periods 52
Figure 19: Distribution of transformed average login times, all conditions & periods 53
Figure 20: Distributions of transformed average login times, all conditions & periods 54
Figure 21: Example images from the house images condition 94
Figure 22: Example images from the face images condition 95
Figure 23: Example images from the object images condition 96
Figure 24: Screengrab from the photo blog website 97
Figure 25: Screengrab from the message board website 98
Figure 26: Screengrab from the blog website 99

List of Appendices

Appendix A. Participant Consent Form 81
Appendix B. Participant Information Forms 83
Appendix C. Session One Post-Task Questionnaire 87
Appendix D. Session Two Post-Task Questionnaire 90
Appendix E. Debriefing Form 91
Appendix F. Participant Reminder Email 92
Appendix G. Recruitment Poster 93
Appendix H. Image Examples 94
Appendix I. Website Examples 97

Introduction

The effectiveness of any computer security system depends on proper use. Security experts will sometimes refer to people as "the weakest link in the chain" of system security (Sasse & Flechais, 2005). Even attackers have come to this realization, with many preferring to employ strategies that target user carelessness or ignorance. While it is true that security systems are often rendered ineffective because users fail to use them properly, this failure can also be seen from the perspective that improper system design is to blame. One of the reasons that current security systems suffer is because they fail to incorporate human factors knowledge in their design. As humans, we have cognitive limitations that restrict and define the potential for our interaction with computers. System designs that fail to take these human factors into account will inevitably lead to failure. However, potentially more constructive than a review of the limitations of our cognitive capacity is to consider and leverage its strengths.

Authentication is the process of ensuring that a user has the right to access resources or services. Today, the most commonly used authentication systems involve the use of text passwords. A usable password must be easy to remember. However, a secure password must be hard to guess. Thus the challenge presented to usable security is to create authentication schemes that employ "passwords" that are easy for legitimate users to remember and use, but difficult for attackers to obtain or guess. Unfortunately, it just so happens that what is typically memorable is also quite guessable. This problem illustrates a trade-off between usability and security. As we increase either security or usability, the other tends to suffer. For example, consider the situation in which a system promotes security by demanding that passwords consist of long strings of random letters, numbers, and special characters. While it would be difficult to guess such passwords, remembering them would be a challenge for most users, and remembering several of

them would be nearly impossible without writing them down. Indeed, a large part of the problem with password authentication comes from the cognitive limitations of human memory.

To make matters worse, typical web users have a growing number of usernames and passwords. Users are becoming overwhelmed with the task of remembering more and more login information. It has come to the point where some web users will simply refuse to use a new website if they deem its authentication system unnecessary. Indeed, users are asked to try and remember too many passwords — a difficult position that leads to all kinds of undesirable effects and behaviours. For example, people resort to actions that compromise the very purpose of such security systems, like re-using passwords, writing their passwords down, or sharing them with friends. Users are often forgetting and resetting their passwords — a process which is costly and irritating for both the user and service providers (Mandylion, 2010). Bill Gates, co-founder and chairman of Microsoft, called for an end to passwords at the 2006 RSA Conference on information security: "I don't pretend that we are going to move away from passwords overnight, but over three or four years, this change can and should happen," he said (Fried & Evers, 2006). The need to prove one's identity by authenticating is not going anywhere. There will always be reasons to restrict access to resources, and to ensure the accountability of those who do access resources. Thus, alternative authentication schemes are being developed and explored.

When exploring more usable alternatives to the ubiquitous text password, we might consider some known cognitive propensities. For example, we know that pictures are more easily remembered than words (Madigan, 1983; Nelson *et al.,* 1977; Paivio *et al.,* 1968). Indeed, human ability regarding visual memory is known to be quite strong. Furthermore, there may be a particular aspect of visual memory, like facial recognition, that is especially promising for use in

authentication schemes. This is the idea behind *PassFaces*, a commercially available authentication system that has already been implemented by a number of large websites (PassFaces, 2010). In PassFaces, users authenticate by correctly selecting pre-chosen faces from random sets of distractors, as shown in Figure 1. The pre-chosen set of faces is what makes up the user's "password", which is called a *graphical password*, since it is based on images rather than text.

Figure 1. **Sample of the PassFaces login screen.**

The purpose of this study is to further explore such graphical password schemes. More specifically, the study examines the nature of visual memory for different image types, and explores the usability potential of graphical passwords at an acceptable security level.

Graphical Passwords

With graphical passwords, users interact with images in order to authenticate. As such, a user "password" is not a word at all. Instead, the "password" arises out of the interaction that takes place between the user and the image(s). De Angeli *et al.* ((2005) suggest that there are three types of such graphical password schemes: *drawmetric* schemes, *locimetric* schemes, and *cognometric* schemes.

With drawmetric schemes, such as Draw-a-Secret (Jermyn *et al.,* 1999), users must recreate figures or shapes by drawing them in order to authenticate, as shown in Figure 2. One of the problems with this scheme is simply that it can be a challenge for users to recreate their drawings accurately enough for the system to recognize them. The other problem arises out of the fact that the user task is one of pure *recall* (De Angeli *et al.,* 2005). When a user comes to draw their "password", they are doing so on a more or less blank space. As a result, they must recall their drawing from memory without cueing and instead of recognizing it. A case might be made that the drawing space itself acts as a sort of cue for users. However, since the same cue is presented to all users (and attackers) for all their accounts, it serves as more of a mild contextual cue than a password-specific one. The difference between recall and recognition will be discussed in a later section, but human performance is typically higher for tasks of recognition.

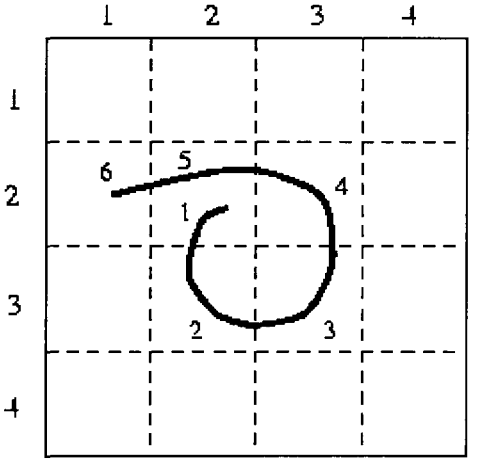


Figure 2. **Sample of the Draw-a-Secret (DAS) login screen, showing a user drawing.**

Locimetric schemes present an image that acts as a cue, and users must click on particular regions of the image. Therefore, the proper regions comprise the "password". Originally

developed by Wiedenbeck *et al.* (2005), *PassPoints* is the most studied such scheme, and it is shown in Figure 3.

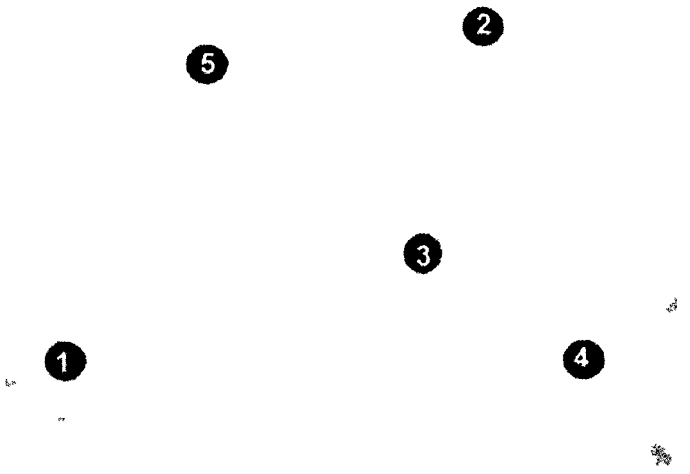


Figure 3. **A sample PassPoints image, showing user-selected click points.**

The problem with locimetric schemes is that there is a potential for the user to choose very obvious points on the image. Security analyses conducted on different versions of the PassPoints scheme have found that users' selection of points on a given image is not random, but instead follows patterns and is focused around so called "hotspots" (Dirik *et al.,* 2007; Thorpe & van Oorschot, 2007). The predictability of user selection is a security concern for these schemes. To address the problem, a modification of the scheme called *Persuasive Cued Click-Points* (PCCP) was developed and tested (Chiasson, Forget, Biddle, & Van Oorschot, 2008). With PCCP, users are guided to choose more random passwords. This is done by restricting user choice to a small, random area within the larger image. The user has the option to request a different area if they are unable to choose a memorable point within the area given to them. In a lab-based user study of this method, it was found that the points selected by participants did not follow any patterns or fall on known hotspots, and that login success rates approximated those of

traditional locimetric schemes (Chiasson *et al.,* 2008). One of the main problems remaining with PCCP is one that affects all graphical password systems — that of "shoulder-surfing". Even a coincidental passerby might accidentally see where a user is clicking, and those actually trying to shoulder-surf are potentially much more dangerous.

In cognometric schemes, users are presented with sets of images and are required to correctly recognize and select a particular image in each set. The "password" is the series of particular images that the user must select. There are many variations on this scheme, but the one most used and studied is called *PassFaces*. In *PassFaces*, users are given a set of faces to remember. In order to log in, the user must recognize and select each face, as it appears randomly among a grid of distractor faces. A problem with this scheme is that in versions where users are allowed to choose the faces that make up their password, significant biases are found in user selection. Specifically, users predictably choose attractive faces of their own race (Davis *et al.,* 2004). From a security standpoint, this vastly increases the chances that attackers will correctly guess user passwords. On the other hand, assigning random images to the user may have negative implications for memorability. Users are more likely to remember passwords that they have chosen than ones that have been assigned to them (Renaud, 2009). The decision whether or not to allow users to select their own images illustrates yet another example of the typical trade-off between security and usability. The challenge is to find the right balance between the two objectives. Since recognition-based graphical password systems have been found highly usable to begin with (Brostoff & Sasse, 2000; Dhamija & Perrig, 2000; Valentine, 1999), the gain in security by assigning images to the user is hypothetically deemed worthwhile. Importantly, the cognometric scheme in the present study involves both assigned passwords and

an increased security level by including more rows, columns, and panels than is typically seen in graphical schemes such as PassFaces.

Related Theory

Visual Memory

The human mind possesses an exceptional ability to remember images — an ability that has been studied for some time (Bower *et al.,* 1975; Calkins, 1898; Paivio *et al.,* 1968; Shepard, 1967; Standing, 1973). Moreover, studies have demonstrated that pictures are recalled and recognized more easily than words (Bower, 1972; Paivio, 1969). Despite this knowledge, the predominant authentication scheme hinges upon the free recall of text passwords.

A study by Bousfield *et al.* (1957) showed that recall of words has a positive relationship with the amount of "signs" shown with them during original exposure. Participants were shown only words in condition A, words accompanied by corresponding black and white images in condition B, and words accompanied by coloured images in condition C. Participants were then asked to recall as many of the words as possible. Results confirmed the hypothesis that in terms of ability to promote recall, the experimental conditions varied such that $A < B < C$. Naturally, pictures carry more information than a singular word used to describe them. It is this inherent abundance of information conveyed by an image that is responsible for its increased memorability. Dual-coding theory (Paivio, 1968, 1969, 1973) argues that memory of images is stronger than memory of words because images are more likely than words to be processed *both* visually and verbally. Similarly, Craik and Lockhart (1972) proposed the existence of a levels-of-processing effect, whereby the method and depth of processing affects how an experience is stored in memory. Since images present more information to process, and the information may

be semantically rich, images are likely to be encoded in more ways, which results in increased availability or access to them.

Images of faces are a good example of the information richness available when looking at an image. Bruce and Young (1986) identify several *types* (or codes) of information that are encoded in memory by viewing a picture of someone's face. *Pictorial code* gives a basic description of the image, while *structural code* has to do with the specific view and angle of the face. *Name code* is only active where familiar faces are being viewed, but *expression code* will always transcribe the meaning of facial expressions and posture. *Visually derived semantic code* has to do with apparent features like race, age, and gender. *Identity-specific code* is all the known information about the person that is not apparent in the image of their face – occupation, for example. Finally, *facial speech code* is said to record whether or not the face is saying something, and what it might be saying. In summary, these codes contain a great deal of semantic information to be processed, which shows the potential importance of the levels-of-processing effect. Increased memorability might be due to the multiple semantic coding that takes place when looking at a face.

Recognition vs. Recall

When comparing traditional text passwords to *cognometric* graphical passwords (the type used in our study), it is important to distinguish between two memory processes — recall and recognition. Recall takes place when one thinks back in time and brings to mind information of which one was previously aware. The most distinct recall task, *free recall*, takes place when one is faced with a blank space and asked to fill it in — not unlike entering a password into a text box. Cued recall is slightly different in that the subject is given a hint or clue that assists

memory. Recognition occurs when one correctly identifies someone or something that they already know, when it is presented to them at a later time.

Our ability to recognize is generally found to be superior to our ability to recall it (Kausler, 1974). This point is commonly illustrated by examples from everyday life. For example, multiple choice questions are frequently easier than essay questions because the correct answer is available for recognition. There is also the tendency to recognize a person's face, but forget their name.

While the superiority of recognition is generally acknowledged, there have been a few studies that show little difference, or even the reverse effect. In particular, it appears possible that serial effects allow people to recall items in order very effectively — more so than recognizing them individually (Dale, 1966; Watkins, 1974).

Face Recognition

The PassFaces graphical password scheme is based on the idea that humans have a special ability for the recognition of faces. There is an increasing amount of evidence that there may be regions of the brain dedicated to facial recognition and processing (Farah, 1996; Haxby, Hoffman, & Gobbini, 2000; Minnebusch, Suchan, Koster, & Daum, 2009).

Prosopagnosia (face blindness) is a disorder whereby the ability to recognize people's faces is impaired. Prosopagnostics must rely on other cues to recognize people (voice, clothing style, etc.). While prosopagnostics have been known to occasionally struggle with object recognition, they struggle much more with faces. *Visual agnosia* is a more general condition where sufferers are unable to make use or sense of normal visual stimuli. However, facial

recognition is preserved in many cases of visual agnosia. This suggests that facial recognition is functionally different than recognition of other visual stimuli (Farah, 1996).

Functional MRI studies have shown that a region in the fusiform gyrus known as the "fusiform face area" (FFA) is activated more strongly when viewing faces as compared to other stimuli. Bilateral FFA activation takes place in both face and object processing. However, when processing faces, right hemisphere FFA activation is a relatively higher correlate. Also active during face processing is the superior temporal sulcus (STS), which appears to process dynamic aspects of facial information, like expression or gaze. There is also an important area in the occipital region termed the "occipital face area" (OFA), which is sensitive to physical features of face stimuli. Minnebusch *et al.* (2009) outline the consensus that all three of these regions are recruited bilaterally (with some right hemisphere bias) during facial processing. More specifically, evidence from their study indicates that both bilateral OFA activation and left FFA activation show significant covariation with activation in the right FFA. When the integrity of this network is disrupted by malfunction in one of the above mentioned areas (as in prosopagnosia), facial processing is impaired.

Haxby *et al.* (2000) distinguish between recognition of the *invariant aspects* of faces that are responsible for identifying individuals, and the *changeable aspects* of faces such as expression, lip movement, and eye gaze. The invariant aspects are handled by the face-responsive region in the fusiform face area (FFA). The changeable aspects are dealt with by the face-responsive region in the superior temporal sulcus (STS). These occipitotemporal regions in the extrastriate visual cortex act in concert with neural systems used for other cognitive functions to obtain meaning from faces. This viewpoint is broadly consistent with the work cited above, and all this research indicates specific neurological support for face recognition.

There is also evidence from beyond the realm of neuropsychology that facial recognition is distinct from recognition of other visual stimuli. Farah, Wilson, Drain, and Tanaka (1998) conducted several studies involving the retroactive interference of different types of stimuli in same-different matching tasks. Their main findings provide strong operational evidence for the unique holistic processing of faces. The holistic nature of facial recognition refers to the idea that faces are interpreted using a "Gestalt" impression resulting from the overall set of facial features.

Other researchers have studied the vertical inversion of visual stimuli, which disrupts the recognition of faces more than the recognition of other objects. This effect, known as the *face inversion effect*, is often used to support the idea that face recognition is a dedicated process that is different from general object recognition (Zhao, Chellappa, Phillips, & Rosenfeld, 2003).

From infancy, we are innately drawn toward faces (Johnson, Dziurawiec, Ellis, & Morton, 1991), and over time we all become experts at recognizing and interpreting faces. Diamond and Carey (1986) present data that suggests a role of expertise in facial recognition. They distinguish between what they called *first order relational information* and *second order relational information*. First order relational information pertains to the relative spatial arrangement of parts of an image. Second order relational information compares the general spatial arrangement to that of the prototypical arrangement for images of that type. Diamond and Carey suggested that the use of second order information and prototypes is not unique to face recognition, but instead underlies any "expert" recognition of images with prototypical arrangements. They showed, for example, that dog experts suffer from an inversion effect similar to the facial inversion effect when viewing inverted images of dogs, while non-experts do not. Diamond and Carey do not doubt the neurobiological support for face recognition, but show that extensive expertise may also be involved the high performance of face recognition that is

observable. It appears that we are all innately predisposed to be experts at face recognition, but we then spend years developing that expertise.

It is well documented that individuals show better recognition for faces of people from their own race (Sporer, 2001). In his authoritative paper, Sporer compares several meta-analyses (including Meissner and Brigham, 2001, for example) on this topic, known as the own-race bias (ORB), or the cross-race recognition deficit. He concludes that "as the results of these meta-analyses suggest, the differential recognition of members of another ethnic group can be regarded as a robust phenomenon" (p. 48). The author explores a few key theories as to why this occurs. For one, differential experience leads to differences in the amount of contact an individual has with members of different racial groups. Perhaps unsurprising is that the quality of this contact is more important than the quantity. Once again of importance is the levels-of-processing approach by Craik and Lockhart (1972), as individuals may process images of those from the "out-group" at a more shallow level than those of the "in-group". Recall that depth of processing is said to have implications for later memory. Interestingly, the typical cognitive disregard for members of the out-group can be mitigated by the perceived social utility of that out-group (Sporer, 2001). Importantly, Chiroro, Tredoux, Radaelli, and Meissner (2008) make the distinction between differences of race and differences of ethno-geographic origin. They present evidence that race categories (like black or white) are not perceptually homogeneous, and that within-race face variation (based on geographic region) can occur. Predictably, their study showed that both black and white South-African participants better recognized members of their own race. However, both racial groups also had better recognition for members of their own race that are from their own geographic region than members of their own race from a different geographic region (the United States). For this reason, the authors suggest "it is perhaps time that

the concepts of *race* and, specifically, *own-race bias* be retired from this literature" (p. 1091). Instead, they argue in favor of referring to the phenomenon as the *in-group face recognition advantage*.

If a cognometric graphical authentication system such as PassFaces randomly assigns users with a series of faces from a variety of ethnic origins, the number of faces in a password that are of the same race of a given user can vary. This may have differential effects on the memorability a given password. For example, a visible minority may be disadvantaged when trying to remember a series of faces from a database that is (perhaps unsurprisingly) comprised mostly of members of the majority. Initially, this may appear to be a disadvantage of using face images for cognometric graphical authentication schemes. However, in the same way that perceptual expertise is said to cause an imbalance in processing that leads to differential recognition ability for face images of a given ethno-geographic type, so too does it affect other imagery. For example, consider a cognometric scheme that uses images of random objects. The objects would inevitably have a certain cultural bias to them, and may not be processed in the same way by those from cultures where the featured objects are less common. The ideal scheme presents the user with subject matter of which they are most familiar. One way this could be accomplished is by allowing the user to choose from a selection of image types, advising them to choose the type of image with which they are most familiar.

Password Space

The trade-off between usability and security suggests that as we increase either security or usability, the other tends to suffer. A critical factor in the usability of any password scheme is the memorability of the password. To study possible differences in memorability, it is necessary

to be consistent in the level of security provided. Otherwise, we are comparing apples to oranges. In computer security, the strength of a password scheme is described in terms of its password space: the number of possible passwords. It is important to distinguish between the *theoretical password space* and the *effective password space*. The theoretical password space is the number of all mathematically possible passwords. The effective password space is the number of all passwords likely to be used in practice. The larger the password space, the more difficult it is for an attacker to guess correctly. Of course, attackers may also be able to capture passwords by means of covert observation or recording software, but that threat model is not addressed in the present study.

The theoretical space of a text password system involves the number of possible characters and the length of the passwords. For example, imagine a text password system where the only requirement is that the password be at least six letters long, and is not case sensitive. The number of possible passwords is therefore $26^6 \approx 300$ million. The space from which that password is drawn is much smaller than that of a password required to be at least eight characters long and include numbers and special characters. The number of distinct characters on a standard US keyboard is 95, thus the number of possible passwords would be $95^8 \approx 6.6$ quadrillion. Table 1 shows the range of possible password criteria and the resulting theoretical password spaces. Because of the exponentially varying sizes of password spaces, they are typically described logarithmically in terms of base two logarithms. For example, $\log_2 26^6 \approx 28$, referred to as 28 bits, and $\log_2 95^8$ translates to 53 bits.

Table 1. Text password criteria and theoretical space.

Description	Number of chars.	Length	Space	Bits
PIN	10	4	1.00E+04	13
lowercase	26	6	3.09E+08	28
lowercase	26	8	2.09E+11	38
mixed case	52	6	1.98E+10	34
mixed case	52	8	5.35E+13	46
alphanumeric	62	6	5.68E+10	36
alphanumeric	62	8	2.18E+14	48
full keyboard	95	6	7.35E+11	39
full keyboard	95	8	6.63E+15	53

A secure text password system will ask users to create passwords that are of at least a minimum length, and that include letters, numbers and special characters. While this does result in more secure passwords, users will typically create a password that just meets the minimum requirements. Moreover, they tend to do it in a predictable way — for example, by adding one number and one special character at the end of an otherwise guessable password. This is just one example of how "users tend to circumvent restrictions that they find tedious" (Yan, Blackwell, Anderson, & Grant, 2004). It is important to understand that the resulting *effective* password space is therefore smaller than the *theoretical* password space. This is because users choose text passwords that are far less random than they could be (Feldmeier & Karn, 1990; Klein, 1990; Morris & Thompson, 1979; Wu, 1999).

For recognition based, cognometric password schemes (like PassFaces), the theoretical password space depends on the number of rows, columns, and panels in a given system. The number of rows and columns determines the number of faces in each panel, and the number of panels determines the number of targets. Table 2 shows these criteria and the resulting theoretical password spaces. For example, in a system that used five rows, five columns, and six

panels, the theoretical password space is $\log_2 (5 \times 5)^6 \approx 28$ bits. Note that this scheme therefore has the same theoretical space as a text password that is six letters long and single case. In order to achieve the theoretical password space afforded by a mixed case text password that is eight characters long and requires the use of numbers and special characters (53 bits), a cognometric scheme would require twelve rows, eight columns, and eight panels.

Table 2. Cognometric password criteria and resulting theoretical space.

Rows	Columns	Panels	Space	Bits
3	3	4	6.56E+03	13
5	5	6	2.44E+08	28
5	5	8	1.53E+11	37
5	10	5	3.13E+08	28
8	8	5	1.07E+09	30
8	10	5	3.28E+09	32
12	8	8	7.21E+15	53

It is important to note while there is a difference between theoretical and effective password space in typical *text* password schemes, there is no difference in most *cognometric* password schemes. This is because the user typically chooses text passwords, while cognometric password schemes will assign random images in a random order to users in order to avoid the above mentioned predictability of user-chosen passwords in those schemes. The result is a larger actual password space for cognometric password schemes, which is a significant security advantage.

Research Question

Graphical password schemes have recently been introduced as an alternative to text passwords, with the promise of increased memorability. In particular, cognometric graphical password schemes harness the superiority of recognition over free recall. PassFaces further proposes to use images of human faces, which leverages the special human cognitive capacity for face recognition. The PassFaces scheme uses three rows, three columns and four panels, which translates to a password space of 13 bits. When compared to text passwords, this is a relatively small (or less secure) password space. Increasing the security of such a scheme involves adding some combination of rows, columns, and/or panels. The scheme in the present study increased password space to 28 bits by including 5 panels of 26 images, and is a much better comparison to the password space of a typical text password.

In fact, the research question was: How does image type affect the memorability of cognometric authentication schemes when tested at the typically acceptable security level of text password schemes? Also of interest was the usability of cognometric authentication schemes with password space equivalent to that of the typical text password.

Research Study

Research Hypotheses

In order to address our research question, a cognometric authentication scheme with three different image types (see Figure 4 or Appendix H for examples) was implemented. A more detailed description of the image types used in the study is available in the Materials section on page 20.

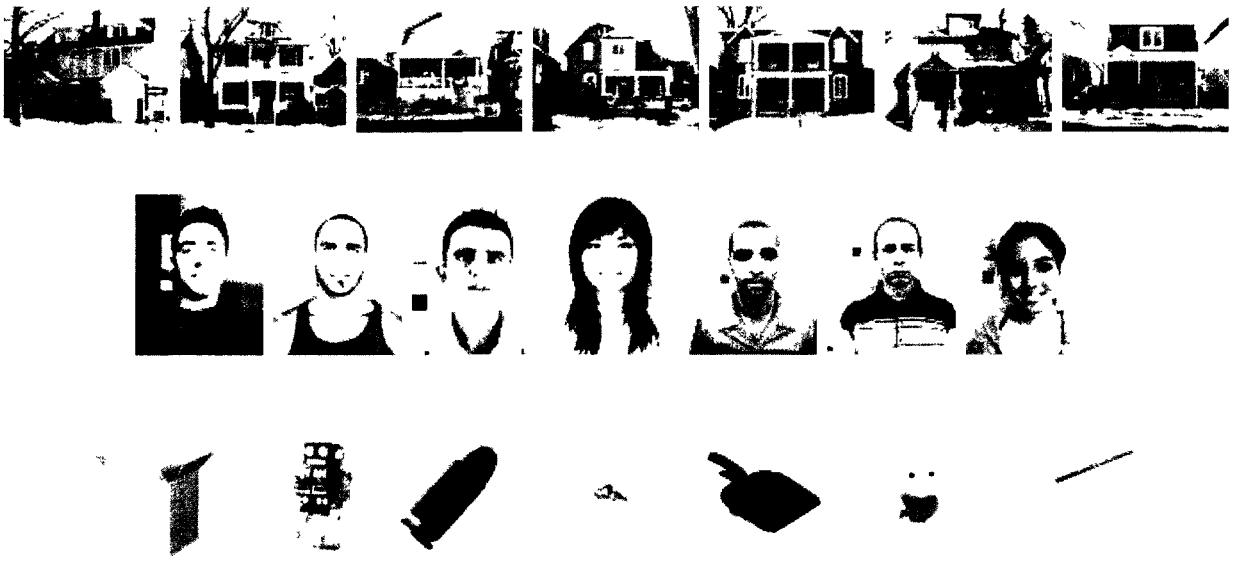


Figure 4. Examples of each image type used in the study. Houses (top), faces (middle), and objects (bottom).

Usability (including memorability) was evaluated by analyzing login times, login errors and login success. Therefore, image type was the independent variable and the dependent variables were to be the number of login errors, login success, and time taken to log in.

$H1_1$: There is a significant difference in the amount of login errors between image types.

$H1_0$: There is no significant difference in the amount of login errors between image types.

$H2_1$: There is a significant difference between image types in login success.

$H2_0$: There is no significant difference between image types in login success.

$H3_1$: There is a significant difference in time taken to log in between image types.

$H3_0$: There is no significant difference in time taken to log in between image types.

There are a variety of reasons to expect such usability differences when altering the type of image used in cognometric graphical password schemes. Since there are arguments that work in favour of more than one type of image used in our study, our hypotheses were non-directional. For example, presented above was evidence to suggest that face images are processed uniquely. It has been assumed by some that this unique processing has implications for memorability. However, unlike the images in the other conditions, the images in the objects condition are vastly *distinct* from one another, both visually and semantically. For example, an image of a freshly sliced lime is vastly different than that of a wooden chair. Hunt (2006) describes a distinct item as "surprising, salient, bizarre, or novel. The subjective experience recruits attention in the form of additional processing that ultimately facilitates memory" (p. 3). It is this *distinctiveness processing*, where differences are found in similar contexts, that leads to beneficial memory effects. The right balance of similarity and difference is vital for distinctiveness processing to take place. At least some level of similarity is necessary to "delineate the episodic context of an item" (p. 22), placing it in a certain category. However, it is the judgment of difference among categorized materials that creates unique psychological representations that lead to better memory performance (Hunt, 2006). The images in the objects condition easily fall into their own category, yet maintain these vast visual and semantic differences. Furthermore, we have a lexicon of associated concepts and interactive protocols for each object that is "called up" when viewing them, which is unique compared to that for other objects. For example, when we think about a banana, what we think about is vastly different than what we think about when thinking about an airplane.

It should be mentioned that, like the selection of face images, the selection of object images may also have cultural biases. This could potentially lead to differences in participant familiarity

(and thus processing) of the object images. Images of houses were included to explore the potential of what Diamond and Carey (1986) refer to as *second order relational information.* House images are like face images in that they vary from a prototypical arrangement of features. Being able to compare participant performance between the face image and house image conditions would allow certain insights into the other advantages of face recognition *not* typically available when looking at images of houses (*expertise* for example). In summary, there is evidence suggesting that different types of processing take place when viewing each image type. This led us to expect differences in various aspects of their usability (number of login errors, login success, and time taken to log in) in the graphical authentication scheme.

Method

Participants

Participants (n = 60) were mainly recruited from the university community. They were contacted through the SONA recruitment system as well as recruited via posters that were displayed around campus (see Appendix G). As such, a large proportion of participants were undergraduate students from Carleton University. The participants were accustomed to entering a username and password to access secure websites on the Internet. The sample was gender-balanced in each condition, and ranged in age from 18 to 43 (M = 21.10, SD = 4.42).

Materials

The recruitment poster, informed consent form, pre-test and post-test questionnaires, email reminder, and debriefing form can be found in the attached appendices.

The images of faces used in our study were obtained with permission from the *Face of Tomorrow* project (Mike, 2010), and included faces of people from diverse ethnic and

geographical origin. However, note that our study did not explore any in-group face recognition advantage. The people in the images used by the commercially available PassFaces authentication scheme (Passfaces, 2010) are all smiling. Smiling faces are more likely to induce positive affect, which is known to have positive effects on human cognitive performance (Ashby *et al.,* 1999). Furthermore, Kirita and Endo (1995) found that when participants were tasked with sorting faces of a variety of expressions into categories, smiling faces were recognized and sorted fastest. The Face of Tomorrow database features a majority of faces that are smiling, but also includes some neutral expressions. Recall the different types of information available when looking at a face identified by Bruce and Young (1986). As *facial expression* is one of those unique types, it could be argued that it may be easier to distinguish among faces when expression is not uniform.

The images in the objects condition were obtained from the stock.xchng website (stock.xchng, 2010). The website terms and conditions permit visitors to use their free images in digital format on websites so long as they are not being used for unlawful, immoral, sales or distribution purposes. The images in the houses condition were based on photographs taken specifically for use in this study. All images used in the graphical password system and on the website were non-offensive to viewers (see Appendix H for examples of images used in the faces, objects and houses conditions, respectively).

Equipment

Participants used personal computers both in the lab and at home (to increase ecological validity). Lab computers were running the Windows XP operating system and the Firefox web browser, and at home participants used which ever was installed on their own computers. All

interaction involved access to three simulated websites with cognometric graphical authentication schemes (see Appendix H).

Procedure

Again, participants were recruited through the SONA system, posters on campus, and by word of mouth. Participants were asked to book two appointments through the SONA system, or through email if they were recruited separately. After being escorted to the study room, participants were given a consent form to read and sign before the study began. They were told that the purpose of the study was to examine the usability of logging into and using websites, and that any problems they had during the study were attributable to the sites, and not to their performance. Participants were told that they would either be paid $20 or receive course credit for their time, and that they would be paid even if they chose to withdraw from the study.

The experiment took place in two sessions held one week apart. Both sessions ran less than one hour in duration. Between sessions, participants logged in with the password system online. Participants were assigned a username and were given a set of six images that made up their password. They were then tasked with logging in, which involved correctly identifying the images in their password among sets of distractors (see Appendix H). Participants were randomly assigned to one of three conditions, each condition using a different type of image. Since participants were assigned their graphical passwords, there was no opportunity for them to share their real passwords with us. Although there were questionnaire items that asked for information about the real passwords that participants used, participants were reminded not to reveal their actual passwords, or any other important personal information.

Session 1

After reading and signing the consent form, participants were reminded that participation included two sessions that were each up to an hour in duration and spread one week apart. They were also reminded that we required them to access our website from home or another remote location in between sessions. Next, participants were asked to fill out a confidential demographics questionnaire. They were then familiarized with the graphical password scheme and how to use it. After a couple of practice trials, they proceeded to regular trials for which data was recorded. Each trial consisted of three stages. In the first stage, participants were presented with the images that were in their password set and were encouraged to spend time memorizing them, after which they would confirm the password by selecting the right images. The second stage consisted of a distraction task, which was a mental rotation task (Peters, 1995). In the third stage, participants were presented with grids of images, and tasked with identifying the images in the grids that were from the password set they were assigned in the first stage. For the sake of external validity, our study aimed to emulate real-life conditions where users have multiple passwords for different systems. To this end, participants were asked to complete three trials in the first session, creating one password for each simulated system that they will access. The first system was a message board, the second was a blog, and the third was a photo-blog. Examples of each simulated system can be seen in Appendix H. When the trials were complete, participants filled out a post-test questionnaire and were debriefed. Participants were also instructed to watch their e-mail over the next few days for messages directing them to log in to the websites.

Online Access

Participants were contacted via e-mail twice during the week after the first session. Each e-mail reminded participants of their second session, and directed them to log in to our websites from the location of their choice, using the same graphical password schemes they were exposed to in the first session. See Appendix F.

Session 2

Participants were asked to return to the lab one week after their first session. During the second session, participants again used the same graphical password scheme that they used in the first session. After their login attempts, they completed a posttest questionnaire about graphical passwords and were debriefed. The debriefing took place in two steps (Part A given after the first session, and Part B after the second session) and consisted of an explanation about the purpose of the study and debriefing forms. Finally, participants received either their $20 honorarium or course credit, and were thanked for their participation.

Analysis Plan

The website that participants used to authenticate logged all activity that took place during each session. Obtained from these logs were the number of real login attempts, the time spent in the login process, and whether each attempt was a success or a failure. From this data the number of errors per trial, the amount of time spent logging in per trial, and the overall success of participants when using the scheme was calculated. Before conducting the suggested parametric statistical tests, the assumptions of each test were checked. In cases where the

assumptions were violated, non-parametric tests that made no such assumptions were used instead.

Hypothesis One

$H1_1$: There is a significant difference in the amount of login errors between image types.

$H1_0$: There is no significant difference in the amount of login errors between image types.

In order to test this hypothesis, an analysis of variance between the mean amount of errors committed for each image type would tell us if they were significantly different from one another. The independent variable would be image type, and the dependent variable would be the average amount of log in errors.

Hypothesis Two

$H2_1$: There is a significant difference between image types in login success.

$H2_0$: There is no significant difference between image types in login success.

In order to test this hypothesis, a chi square analysis would be used. That test allows for the comparison between categorical data sets to determine the likelihood that differences might occur by chance. The independent variable would be image type and the dependent variable would be the categorization of the login attempt as a success or a failure.

Hypothesis Three

$H3_1$: There is a significant difference in time taken to log in between image types.

$H3_0$: There is no significant difference in time taken to log in between image types.

In order to test this hypothesis, an analysis of variance between the average time taken to log in for each image type would tell us if they were significantly different from one another. The independent variable would be image type, and the dependent variable would be the average time taken to log in.

Other Questions

Other researchers have found a significant *effect of gender* in their studies on graphical passwords (Chiasson *et al.,* 2009), which suggested that such an effect may show up in the present study. Since the study took place in four phases (first session, early online access, late online access, second session), the *difference in login success between phases* was also of interest. It was just as easy to imagine participants forgetting their passwords as the week went by, as to think that they would become more successful at logging in due to practice effects.

When a user has multiple passwords of the same kind, there may be memory interference between those passwords. In such cases, memory of one password may interfere with memory of another. This is certainly evident with text passwords, where people commonly enter the password for one website when logging into another (Chiasson *et al.,* 2009). As each participant was assigned three separate graphical passwords, observations of *interference from multiple passwords* were noted and will be mentioned in the Discussion section.

Results

In this section, each hypothesis is tested at an alpha level of .05. In cases where there are multiple comparisons, the Bonferroni method was used to maintain the overall alpha level of .05.

Other relevant results (such as the effect of gender) are also reported. Lastly, we briefly report on feedback from the questionnaires.

Recall that the study took place over the period of a week. At the beginning of the week was the first session, taking place in the lab. During the week, participants received email directing them to access the websites from the location of their choosing. Between six and eight days after the first session, participants returned to the lab for the second lab session. Thus, the data was categorized by the time period that it came from. Specifically, the data was categorized as coming from the first session, from one to two days after the first session, from three to four days after the first session, or from the second lab session. Any activity between day four and the second lab session was ignored. For some analyses, an aggregate variable including data from all time periods was calculated.

Hypothesis One

$H1_1$: There is a significant difference in the amount of login errors between image types.

$H1_0$: There is no significant difference in the amount of login errors between image types.

In the analysis plan, it was proposed that an analysis of variance (ANOVA) on the distributions of errors committed for each image type would be conducted. The independent variable was to be image type, and the dependent variable was to be the number of login errors made by participants. It is important to describe what is meant by the term "login error." For the purposes of this hypothesis test, a login error is defined as a failed login attempt occurring immediately prior to the participant either successfully logging in, or giving up. This measure was important because, although the second hypothesis test explored participant ability to

remember their passwords, it did not capture the amount of *difficulty* involved in doing so. Although this hypothesis test was meant to explore how much participants struggled with their passwords, it should be noted that some were simply more willing to keep trying — a possible confound that will be covered in more detail in the Discussion section.

After the average number of login errors was calculated for each participant, those averages were used to create an average number of login errors committed by participants in each condition. In other words, a mean of means was calculated for each condition. For the hypothesis test, it was decided to use the data from the second lab session. This was decided because by the second lab session, all participants have had some practice using the authentication scheme, and there was a large amount of missing data in time periods prior to it. While participant compliance with login instructions during the week was somewhat lacking, all participants showed up for the second lab session (more on this in the Discussion section). Hypothesis testing began by looking at the descriptive statistics and frequency distributions in order to make sure that the data met the assumptions of the proposed parametric test.

Table 3. Descriptive statistics for login errors across conditions in each time period.

Login Errors		N	Mean	Median	S.D.
Lab Session 1	Houses	20	0.45	0.17	0.54
	Faces	20	0.62	0.33	0.75
	Objects	20	0.52	0.00	0.98
1 to 2 days later	Houses	10	0.85	0.75	1.03
	Faces	12	1.67	1.83	1.36
	Objects	13	0.78	0.33	1.07
3 to 4 days later	Houses	8	0.42	0.00	0.79
	Faces	11	1.32	1.00	2.12
	Objects	12	0.93	0.67	1.55
Lab Session 2	Houses	20	0.68	0.67	0.42
	Faces	20	0.59	0.50	0.63
	Objects	20	0.25	0.00	0.44

As evident in Table 3, mean login errors during the second lab session appeared to be such that participants in the house images condition made the most errors, followed by those in the face images condition, with those in the object images condition appearing to make the fewest errors. Further, participants in the object images condition made fewer login errors than those in the face images condition in every time period. The error rate for participants in the house images condition tended to be the highest, but the results prior to the second lab session are misleading due to missing data, and the reasons for this will be covered in the Discussion section.

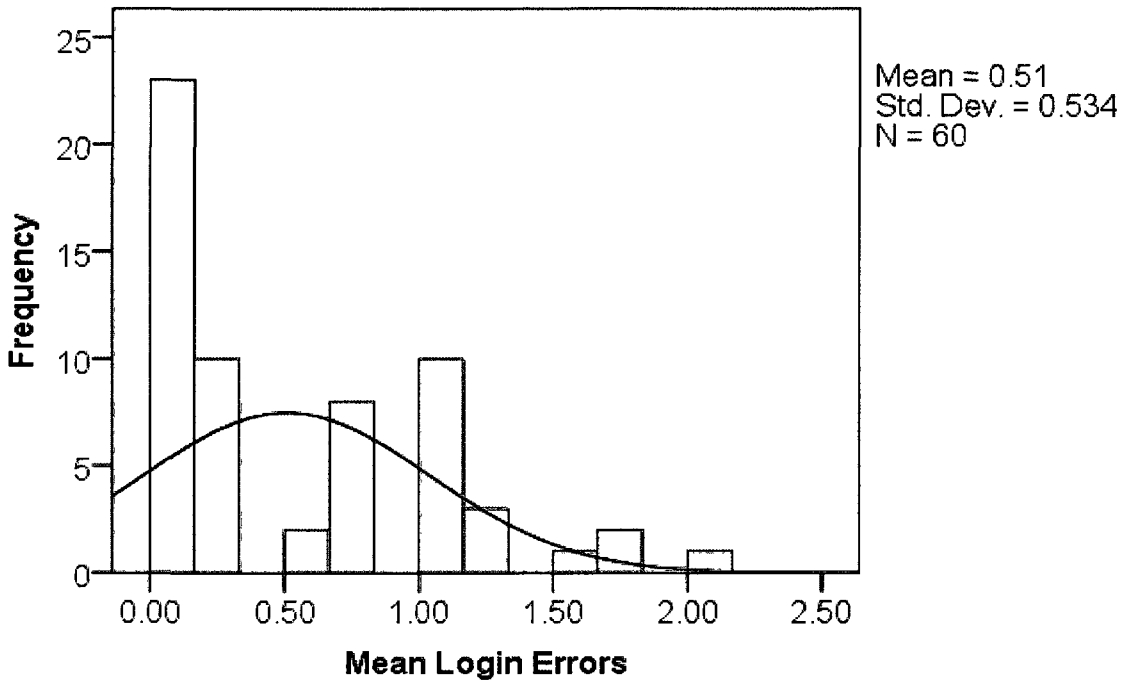


Figure 5. **Frequency distribution (histogram) of mean login errors during the second lab session, including all conditions.**

Next, an analysis was conducted to assess the *statistical significance* of the differences between conditions for average amount of login errors in the second lab session. As evident by the positive skew (skewness = .84) of the distribution in Figure 5, the assumption of normality has been violated. It was proposed in the analysis plan that should the assumptions of ANOVA be violated, we would proceed with non-parametric tests that make no such assumptions. When there are more than two levels in the independent variable, the non-parametric equivalent of ANOVA is the Kruskal-Wallis test, which sorts the data into ranks and then compares the distribution of those ranks. The results of this test indicated that there was a significant difference

between conditions in the distribution of login errors, $\chi_2(2, N = 60) = 10.06, p = .007$, and Figure 6 shows the distribution of login error rates for the three conditions in the study.

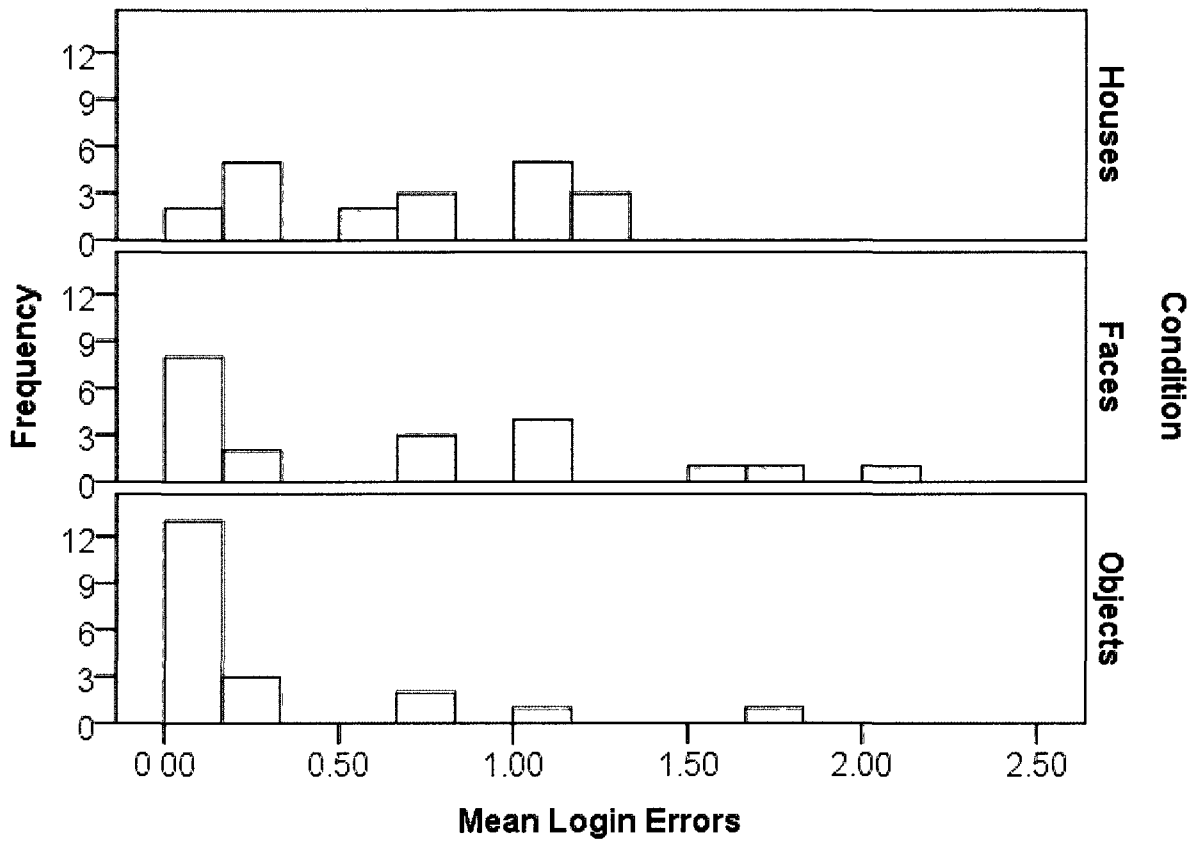


Figure 6. **Frequency distributions (histograms) of mean login errors during the second lab session, separated by condition.**

To uncover exactly which conditions differed from one another, pair-wise non-parametric comparisons were conducted. The non-parametric comparison of central tendency to use when there are two levels in the independent variable is the Mann-Whitney *U* test. The Bonferroni correction was applied to the significance criteria these tests in order to account for multiple comparisons and maintain the overall desired level of significance ($\alpha = .05 / 3 = .017$). The multiple comparisons showed that the only significant difference in login errors during the

second lab session was between the house images condition (*Mdn* = 0.67) and the object images condition (*Mdn* = 0.00), $z = -3.30$, $p = .001$. The difference between the face images condition (*Mdn* = 0.50) and the objects images condition (*Mdn* = 0.00) was considerable, but fell short of the corrected significance threshold, $z = -1.89$, $p = .058$. Lastly, the difference in login errors between those in the faces condition (*Mdn* = 0.50) and those in the houses condition (*Mdn* = 0.67) was not significant, $z = -0.92$, $p = .36$.

In summary, our results suggest that while participants in the object images condition made fewer login errors than those in the face or house image conditions, the only significant difference was between those in the object images condition (the best case) and those in the house images condition (the worse case). There was no support for the notion that the face images condition would have a lower number of login errors. As suggested in the analysis plan, the effect of gender on each dependent variable was also explored. There was no significant difference due to gender in average login errors committed during the second lab session, $z = -0.48$, $p = .63$.

Hypothesis Two

$H2_1$: There is a significant difference in login success between image types.

$H2_0$: There is no significant difference in login success between image types.

In the analysis plan, it was suggested that a chi square goodness of fit test would determine the probability that the observed differences in login success between groups might occur by chance. If the probability was lower than the .05 significance threshold, the null hypothesis (that values would occur in each cell with equal frequency) could be rejected.

However, after collating the data it became apparent that the simplest way to determine *login success* was to look for evidence in each time period that participants had either remembered or forgotten their passwords for each of the accounts they were assigned. Each participant was thus assigned a score between 0 and 3 for the number of passwords they remembered in each time period (0 for no passwords remembered, 3 for three passwords remembered). Note that this variable measured whether participants had *any* success with each of their three passwords. A participant may have been assigned a login success score of 3 because they successfully logged once with each of their passwords in a given time period, but they may also have had many login failures in the same time period — thus the importance of the analysis of login errors in the first hypothesis test.

Thus, the dependent variable was the participant's score, having possible values of 0, 1, 2, or 3, and this forms a ratio variable – a quantitative variable measured from a scale with equal value between possible values that also includes a value of zero representing instances where there is none of that variable. Consequently, the more appropriate test was Analysis Of Variance (ANOVA). Again, the first step was to explore descriptive statistics in Table 4 and frequency distributions in Figures 6, 7, 8, and 9 in order to check the assumptions of the proposed parametric test.

Table 4. Descriptive statistics for login success across conditions in each time period.

Login Success		N	Mean	Median	S.D.
Lab Session 1	Houses	20	2.35	3.00	0.81
	Faces	20	2.65	3.00	0.59
	Objects	20	2.85	3.00	0.37
1 to 2 days later	Houses	10	1.20	1.00	1.32
	Faces	12	1.50	1.00	1.17
	Objects	13	2.54	3.00	0.78
3 to 4 days later	Houses	8	1.63	1.50	1.31
	Faces	11	1.91	3.00	1.20
	Objects	12	1.83	2.00	1.11
Lab Session 2	Houses	20	1.15	1.00	1.31
	Faces	20	1.90	3.00	1.37
	Objects	20	2.35	3.00	0.93

As can be seen in Figures 6, 7, 8, and 9, the distributions of our *login success* data (0-3 passwords remembered) were either heavily skewed toward 0 or 3, or tended toward a bi-modal shape, with most participants either forgetting or remembering all three passwords. As a result, the appropriate statistical analysis method is the Kruskal-Wallis test — the non-parametric equivalent of ANOVA that makes no assumption of normality. Although *login success* consistently varied across conditions such that participants were more successful with object images than faces images, and more successful with face images than house images, the

differences did not cross the threshold for significance in every time period of the study. The first time period in which to look for differences in login success was during the first lab session.

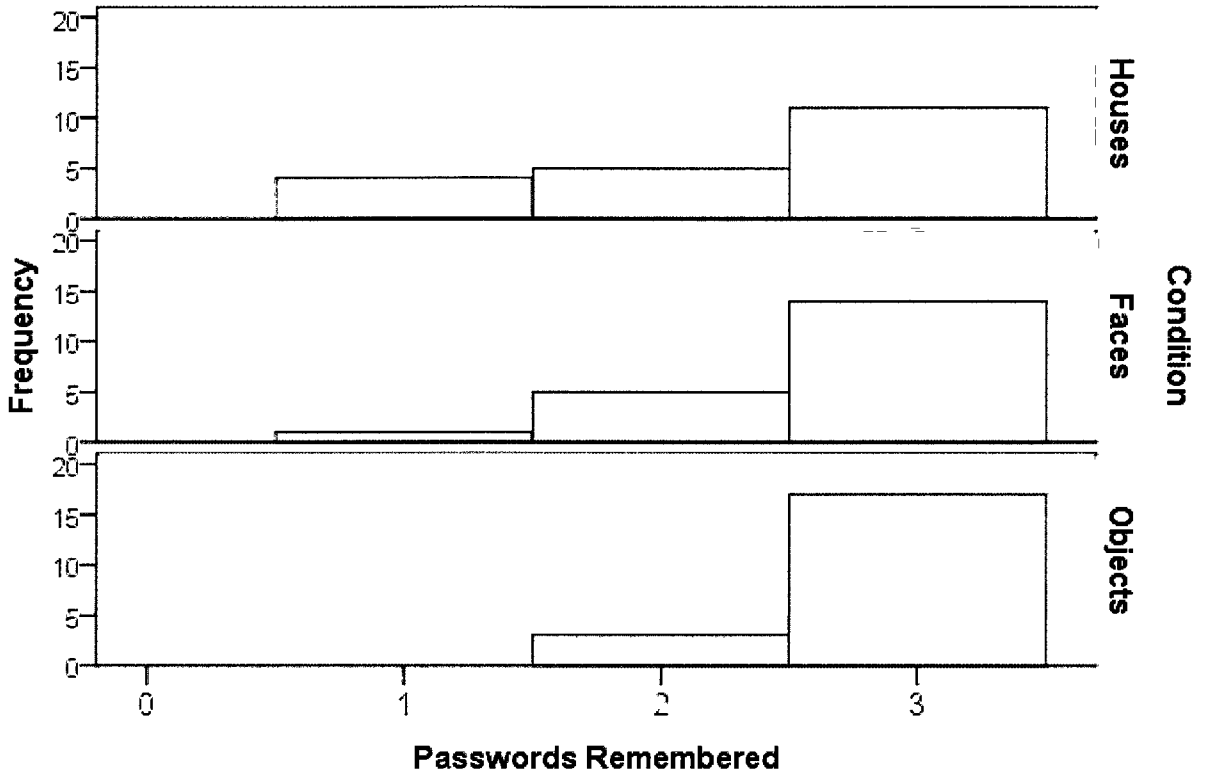


Figure 7. **Frequency distribution of login success scores across conditions during the first lab session.**

Despite the increase in login success apparent in Figure 7 by participants in the object images condition ($M = 2.85$, $Mdn = 3.00$, $SD = .37$) over those in the face images condition ($M = 2.65$, $Mdn = 3.00$, $SD = .59$), and particularly over those in the house images condition ($M = 2.35$, $Mdn = 3.00$, $SD = .81$), the Kruskal-Wallis test revealed that the differences fell short of statistical significance, $\chi^2(2, N = 60) = 5.10, p = .08$. However, bear in mind that the only event in this time period between password creation and login attempts was a very brief distraction

task. The next time period in which to look for differences in login success across conditions was one to two days after the first lab session (see Figure 8).

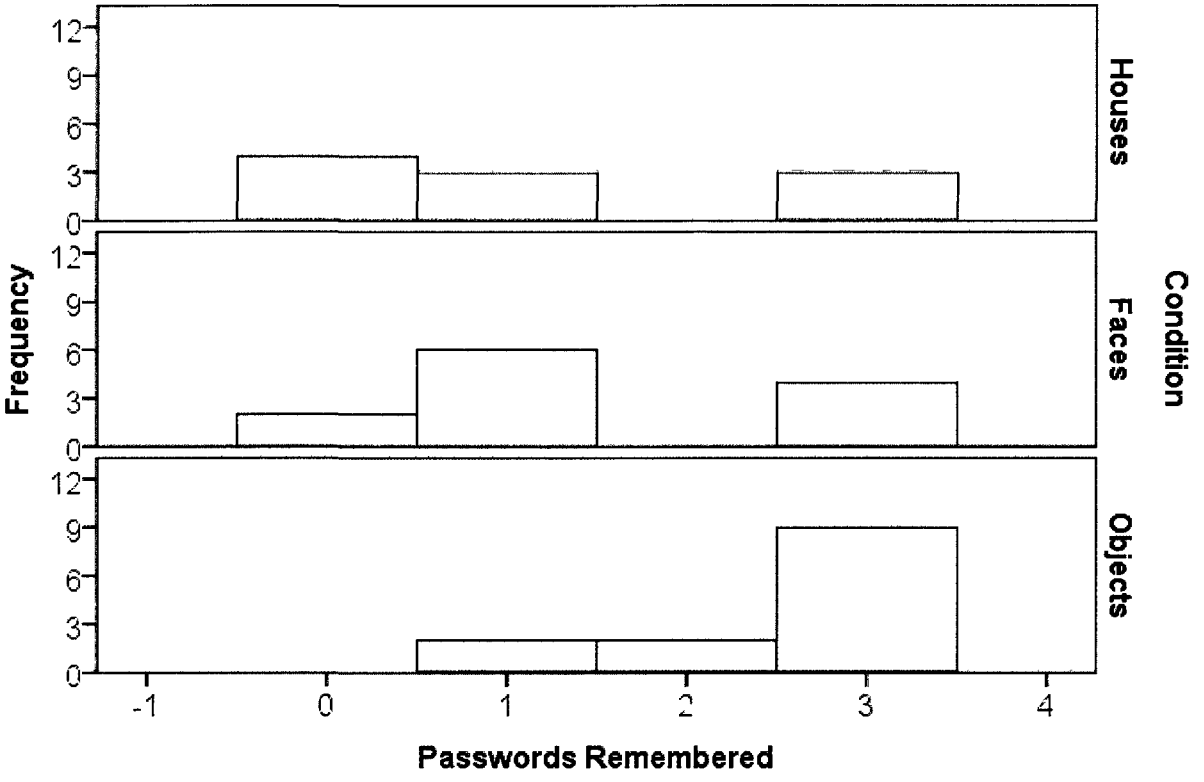


Figure 8. **Frequency distribution of login success scores across conditions one to two days after the first lab session.**

The differences between conditions in login success one to two days after the first lab session were indeed found to be significant, $\chi^2(2, N = 35) = 7.82, p = .020$. Multiple comparisons were then used in order to uncover where the significant differences were. As in the previous hypothesis test, the Mann-Whitney *U* test with the same Bonferroni correction ($\alpha = .05 / 3 = .017$) was used to compare between two levels of the independent variable. Although participants remembered the most passwords in the object images condition ($M = 2.54$, $Mdn = 3.00$, $SD = 0.78$), followed by the face images condition ($M = 1.50$, $Mdn = 1.00$, $SD = 1.17$), with those in the house images condition remembering the fewest passwords ($M = 1.20$, $Mdn = 1.00$, $SD =$

1.32), the only significant difference was between the house images condition and the object images condition, $z = -2.46, p = .014$.

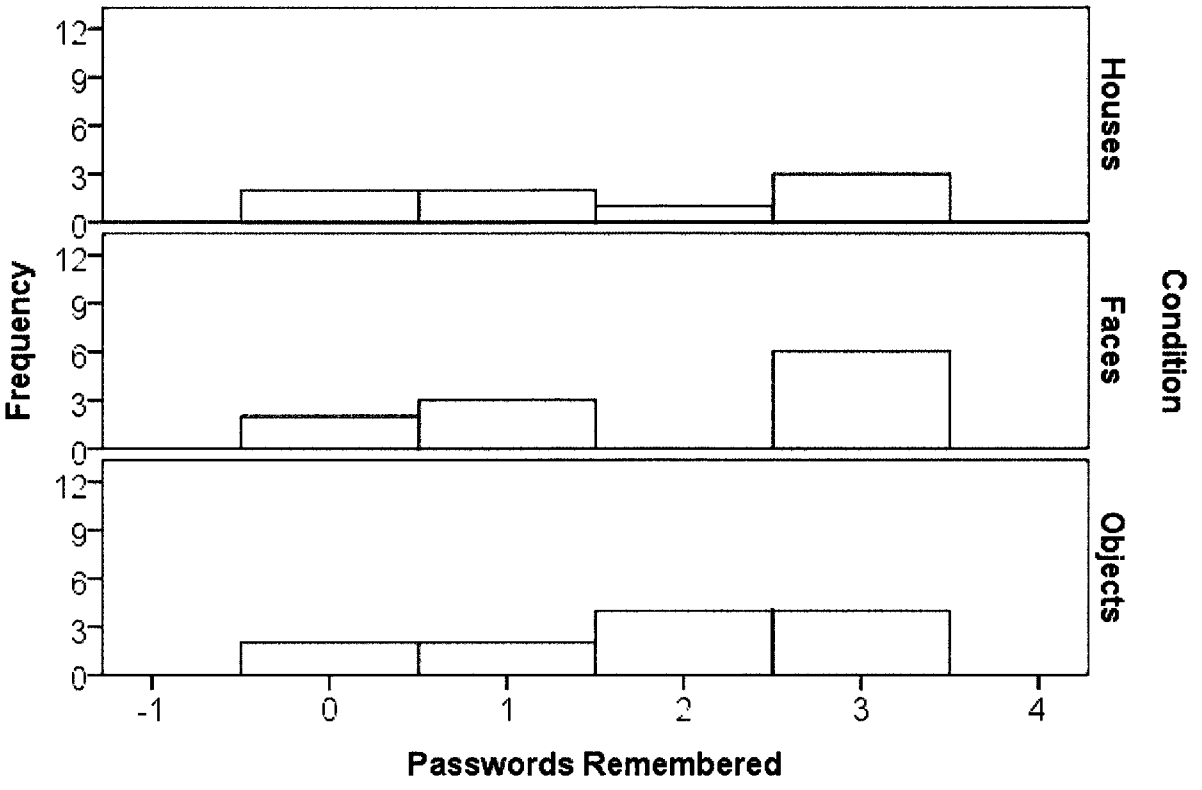


Figure 9. **Frequency distribution of login success scores across conditions three to four days after the first lab session.**

The next time period in which to look for differences across conditions in login success was three to four days after the first lab session. Here there was no significant difference in login success across conditions, $\chi^2(2, N = 31) = 0.32, p = .85$, and there are no apparent differences in the distributions shown in Figure 9. This is in part due to the large amount of missing data for this time period.

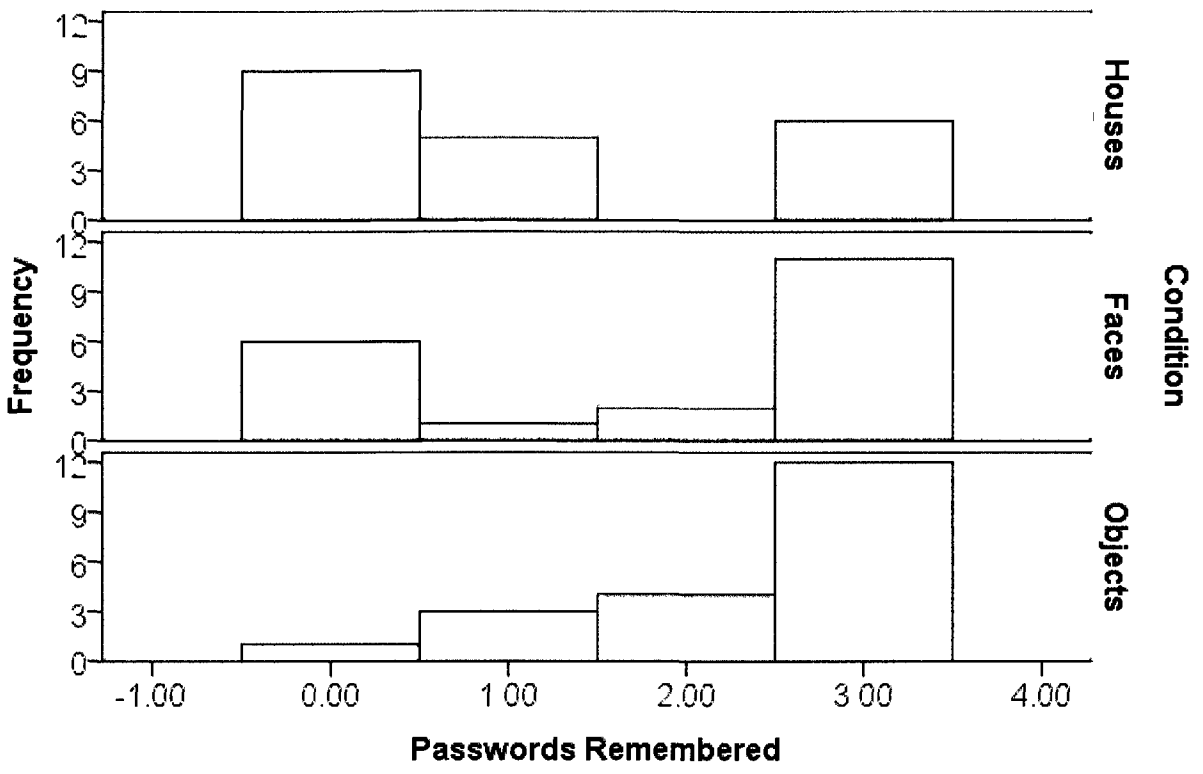


Figure 10. **Frequency distribution of login success scores across conditions during the second lab session.**

Lastly, we also looked for differences in login success across conditions during the second lab session (see Figure 10). Here the differences between conditions in login success were found to be significant $\chi^2(2, N = 60) = 7.62, p = .022$. Again, to uncover the specific nature of those differences, we followed up with Bonferroni-corrected multiple comparisons. Although participants in the object images condition ($M = 2.35$, $Mdn = 3.00$, $SD = 0.93$) appeared to be more successful than those in the face images condition ($M = 1.90$, $Mdn = 3.00$, SD = 1.37) and far more successful than those in the house images condition ($M = 1.15$, $Mdn = 1.00$, $SD = 1.31$), the only significant pairwise comparison was between those in the object images condition and those in the house images condition, $z = -2.85$, $p = .004$.

Our login success findings for this time period are complicated by the fact that some participants had, by this point, reset their passwords. As one of the goals of our study was to maintain some level of ecological validity, we allowed participants who had forgotten their passwords during the week to reset them and learn new ones. As such, the above analysis includes both participants who were attempting to log in with their original passwords, and those who were attempting to log in with new passwords. Naturally, it would seem that those who were using newer passwords would have a distinct advantage in login success. To gain a better understanding of how this may have affected our results, we also compared the number of password resets across conditions.

As the distribution of password resets per participant had a slight positive skew (skewness = .78), with approximately half of all participants deciding not to reset any of their passwords (in some cases, even if they forgot them), it was decided to use the non-parametric Kruskal-Wallis test. Participants in the house images condition ($M = 0.48$, $Mdn = .42$, $SD = .39$) had the most password resets, followed by those in the face images condition ($M = 0.37$, $Mdn = 0.00$, $SD = .47$), with those in the objects condition ($M = 0.22$, $Mdn = 0.00$, $SD = .39$) resetting their passwords the least. However, differences across conditions in average number of password resets per participant fell short of statistical significance, $\chi^2(2, N = 60) = 4.93$, $p = .085$.

The high rate of password resets might be seen as a distinct advantage to those in the house images condition, and to a lesser extent those in the face images condition when attempting to log in at the end of the study. Thus, the inclusion of those who reset their passwords in the analysis of login success during the second lab session may actually amplify the findings for login success during the second lab session. In other words, even though participants in the object images condition reset their passwords the least during the week, they still had the

greatest login success at the end of the study (during the second lab session). The opposite is true for those in the house images condition, with those in the face images condition falling in the middle on both measures. Alternatively, those who had reset their passwords may instead have been at a disadvantage, as they were working with passwords they had used less often, thus having less rehearsal.

Ultimately, the challenge facing participants in this study was to remember their passwords. Of course, failed attempts to log in were almost entirely due to participants forgetting their passwords. Thus, another way to look at login success across conditions is to explore the length of time that participants remembered their passwords across conditions. For each participant and each password, the maximum amount of time between password creation and any successful login was calculated. The average of these times was used to determine *mean memory time for each participant.* Lastly, those means were used to calculate the *mean memory time for passwords in each condition.* The maximum possible time in this measure was the time between the first lab session and the second lab session, which was usually 7 days, or 168 hours. An exploration of the descriptive statistics and frequency distributions would suggest how to go about comparing the means.

Table 5. Mean amount of time (in hours) across conditions that participants remembered their passwords.

Mean Memory Time	Mean	Median	S. D.
Houses	51.57	14.53	67.19
Faces	103.82	124.66	75.96
Objects	134.79	141.08	60.01

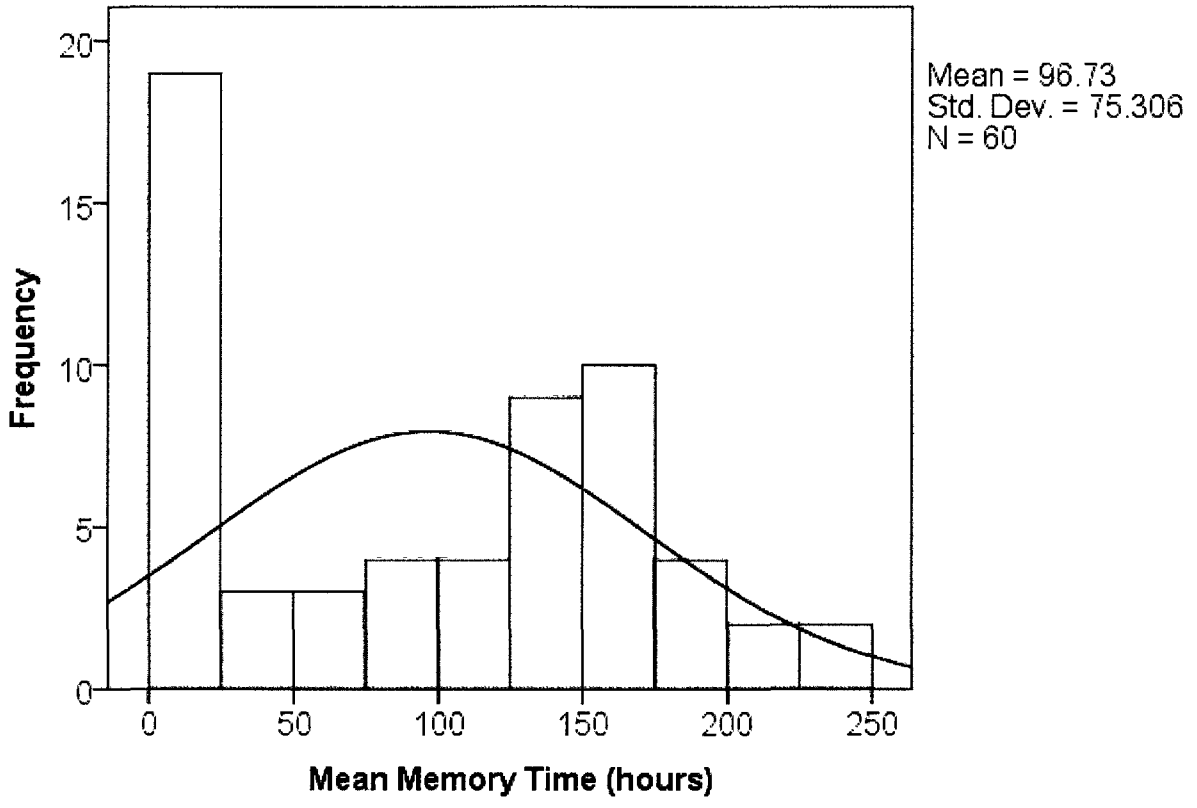


Figure 11. **Frequency distribution of mean memory time, including all conditions.**

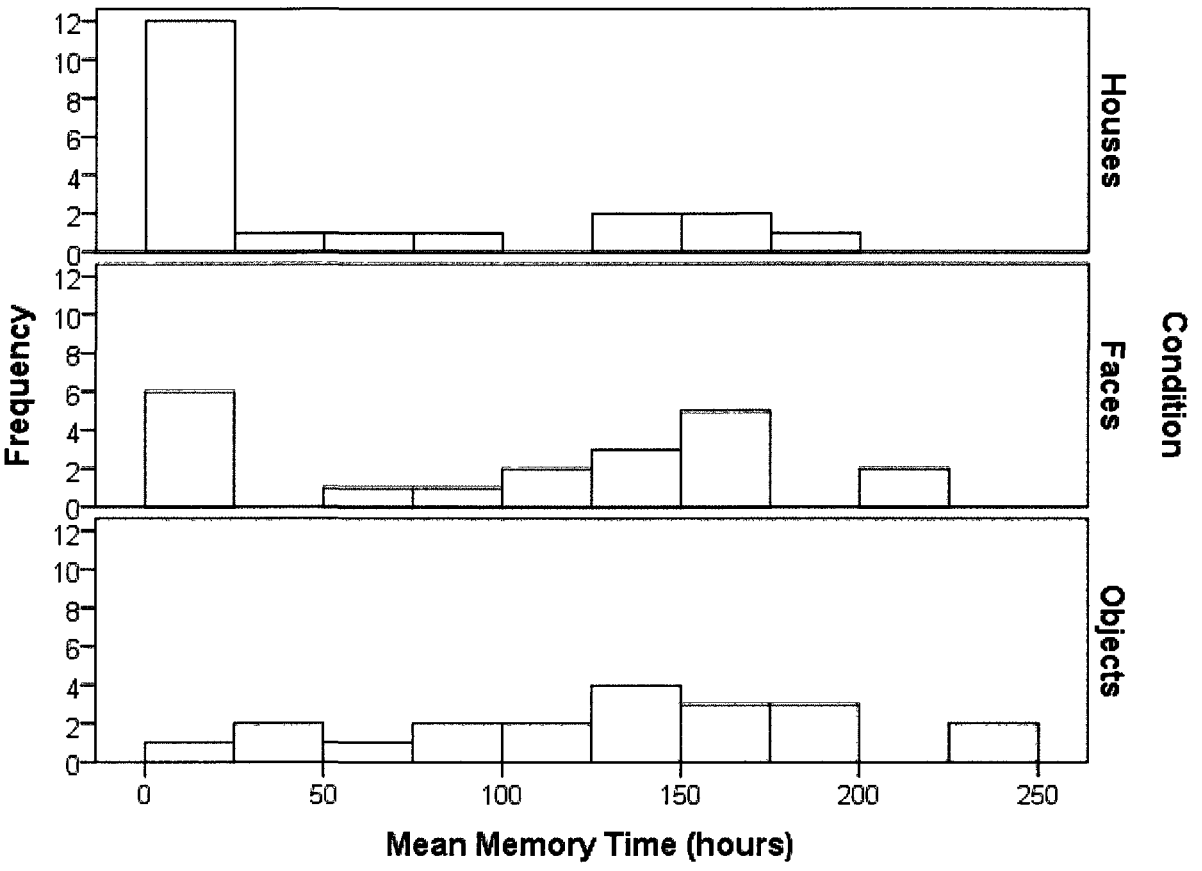


Figure 12. **Frequency distribution of mean memory time across conditions.**

Given the bi-modal shapes of the distributions in Figures 11 and 12, it was decided that the assumption of normality, usually necessary for ANOVA, was not met. As such, we proceeded once again with the non-parametric Kruskal-Wallis test. A significant difference was found in maximum memory time between conditions $\chi^2(2, N = 60) = 12.01, p = .002$. As seen in Table 5, the object images ($M = 134.79$, $Mdn = 141.08$, $SD = 60.01$) were remembered the longest, followed by face images ($M = 103.82$, $Mdn = 124.66$, $SD = 75.96$), and house images ($M = 51.57$, $Mdn = 14.53$, $SD = 69.51$), with large amounts of variance in each condition. Corrected pair-wise comparisons indicated that the only statistically significant difference was between the object image and house image conditions ($z = -3.54$, $p < .001$). Note also that, although it did not

cross the typical threshold for statistical significance, participants in the object images condition remembered their passwords approximately 31 hours longer (on average) than those in the face images condition.

In summary, login success varied consistently across conditions in every time period such that participants in the object images condition remembered the most passwords, followed by those in the faces condition, with those in the house images condition remembering the fewest. However, differences were not statistically significant in every time period and when they were, the only significant pair-wise comparison was between those in the object images condition and those in the house images condition. Also worthy of mention was that participants in the object images condition remembered their passwords the longest and made the fewest password resets, followed by those in the face images condition, with those in the house image condition remembering the passwords for the least amount of time and making the most password resets. Although statistical significance among these findings was often elusive, there was directional consensus from virtually every analysis.

Finally, there was no significant difference in login success across gender in any time period of the study. Differences were not significant during the first lab session ($z = -0.08$, $p = .93$), one to two days after the first lab session ($z = -1.31$, $p = .19$), three to four days after the first lab session ($z = -0.90$, $p = .37$), or during the second lab session ($z = -0.12$, $p = .91$). Differences between gender in mean memory time were also found to be non-significant, $z = -0.70$, $p = .49$.

Hypothesis Three

$H3_1$: There is a significant difference in time taken to log in between image types.

$H3_0$: There is no significant difference in time taken to log in between image types.

The data used to test hypothesis three came from the second lab session, and consisted of the average time taken for each user to successfully login. Including data only from successful attempts leads to a more specific measurement of usability. All data in this comparison comes from login attempts where participants had a good idea of what their password was (their attempt was a success). Thus, differences had less to do with memorability and more to do with the ease with which participants could apply their knowledge in order to authenticate. Importantly, the amount of time it took participants to log in also had certain effects on memorability that will be explained in the Discussion section. Login time data from the end of the study was used for two reasons. First, although participant compliance with e-mail requests to log in from home during the week was fairly low, we had data from all participants during the second lab session (again, more on this in the Discussion section). The second reason login time data from the end of the study was used was because by then, all participants had more practice using the authentication scheme.

As each user belonged to one of the different conditions, mean login times (in seconds) were calculated for each condition and then compared to each other. It was proposed that an ANOVA would be used to look for differences in average time taken to log in across conditions. The hypothesis test began by looking at descriptive statistics and frequency distributions in order to determine whether or not the assumptions of the parametric test (e.g., normality) were met.

Table 6. Descriptive statistics for time (in seconds) taken to log in across conditions in each time period.

Login Time		N	Mean	Median	S.D.
Lab Session 1	Houses	20	48.65	41.50	29.36
	Faces	20	30.73	24.00	21.66
	Objects	20	17.03	16.00	4.55
1 to 2 days later	Houses	6	122.44	129.17	58.10
	Faces	10	65.23	70.50	23.48
	Objects	13	46.97	44.33	13.83
3 to 4 days later	Houses	6	57.89	55.00	15.30
	Faces	9	64.48	67.33	31.84
	Objects	10	47.33	47.00	19.67
Lab Session 2	Houses	11	83.06	65.00	54.75
	Faces	14	41.45	39.67	14.18
	Objects	19	31.03	26.50	16.63

During the second lab session, the average amount of time taken to log in across all conditions was rather high and contained a large amount of variance ($M = 47.35$, $Mdn = 38.58$, $SD = 36.45$). As seen in Table 6, looking more specifically at each condition during the second lab session revealed that those in the object images condition took the least amount of time to log in ($M = 31.03$, $Mdn = 26.50$, SD = 16.63), followed by those in the face images condition ($M =$

41.45, *Mdn* = 39.67, *SD* = 14.18), with those in the house images condition taking the longest to log in (*M* = 83.06, *Mdn* = 65.00, *SD* = 54.75).

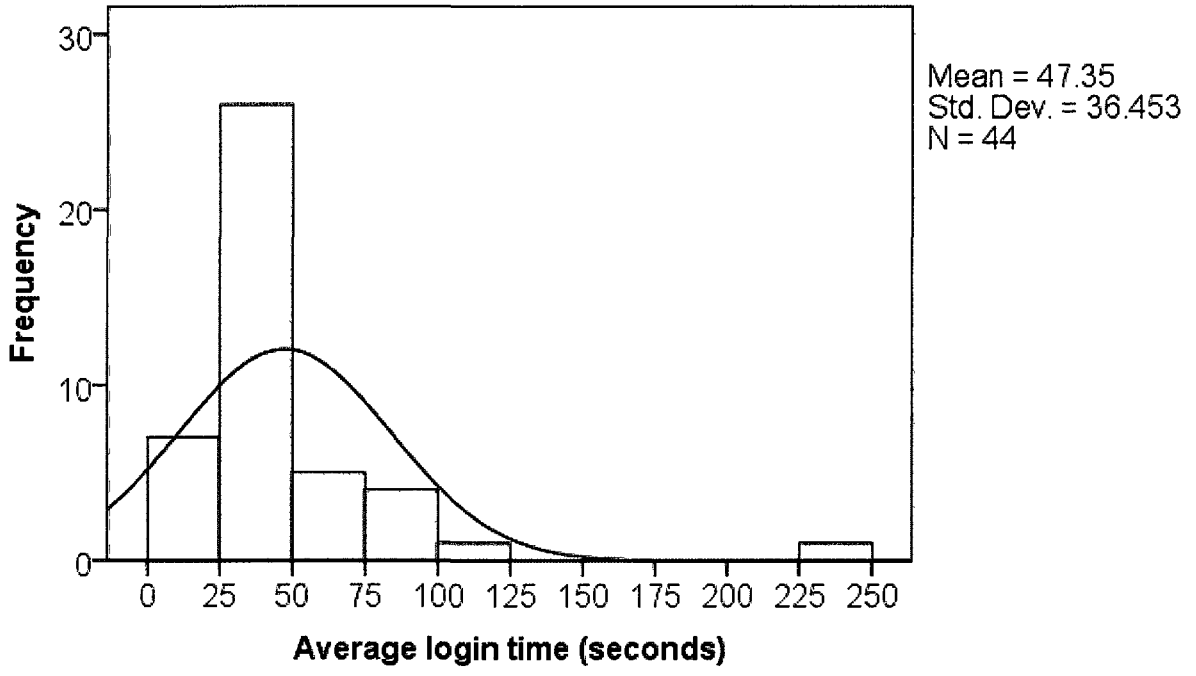


Figure 13. **Frequency distribution (histogram) of average login times during the second lab session, including all conditions.**

The frequency distribution represented in Figure 13 includes data from all conditions, and suggests that the assumptions of ANOVA have been violated. Specifically, the assumption of normality has been violated as both the skewness and kurtosis values lie outside the acceptable range between -2 and 2 (skewness = 3.42, kurtosis = 15.67). Therefore, the non-parametric Kruskal-Wallis test was put to use again. The results of that test were indeed significant $\chi_2(2, N = 44) = 21.30, p < .001$), suggesting a significant difference in time taken to login between conditions.

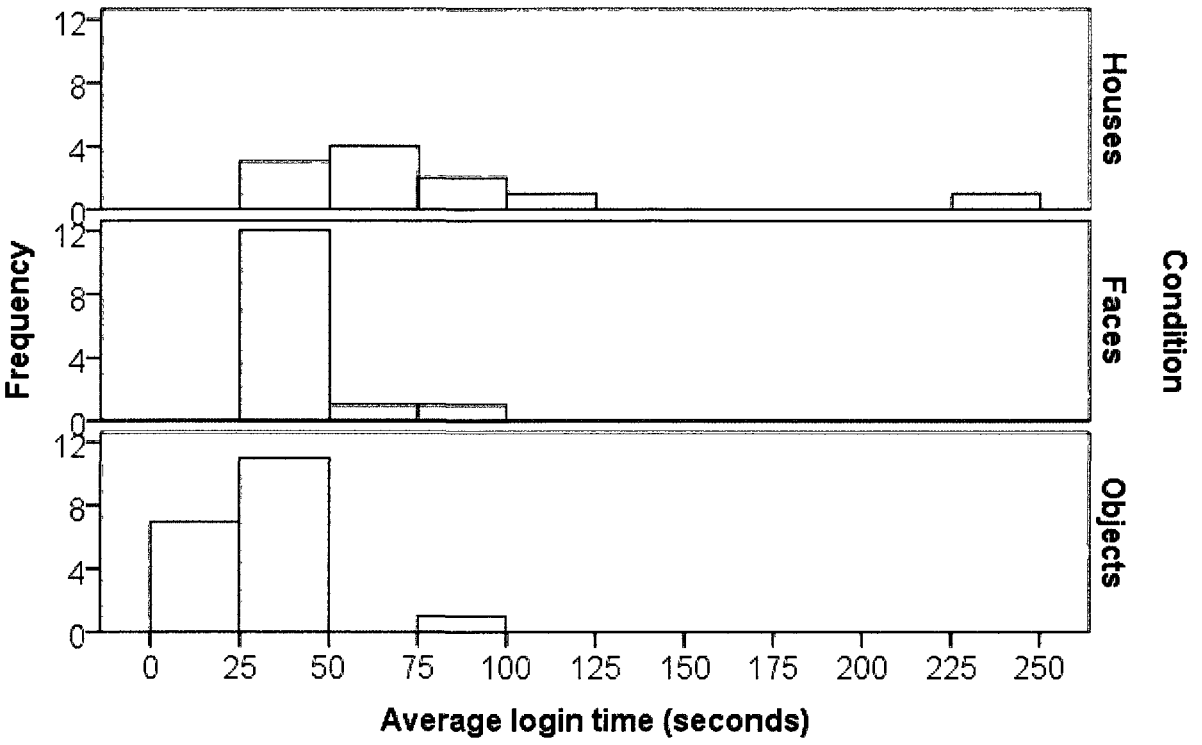


Figure 14. **Frequency distributions (histograms) of average login times across conditions during the second lab session.**

To uncover exactly where (as seen in Figure 14) the significant difference(s) occurred, pair-wise comparisons were then conducted. Once more, the Bonferroni-corrected Mann-Whitney *U* test was applied, finding that those in the object images condition logged in significantly faster than those in the face images condition ($z = -2.53$, $p = .011$), and that those in the face images condition logged in significantly faster than those in the house images condition ($z = -3.12$, $p = .002$). Naturally, it follows that those in the object images condition also logged in significantly faster than those in the house images condition ($z = -4.07$, $p < .001$).

Since non-parametric tests make fewer assumptions about their data, they are sometimes favoured for their wider applicability and greater robustness. However, they typically lack the statistical power of parametric tests when parametric tests are valid. For this reason, a second

analysis method was also used. A base-10 log transformation ($x_1 = \log_{10}(x)$) was used to normalize the data and thus better satisfy the assumptions of ANOVA.

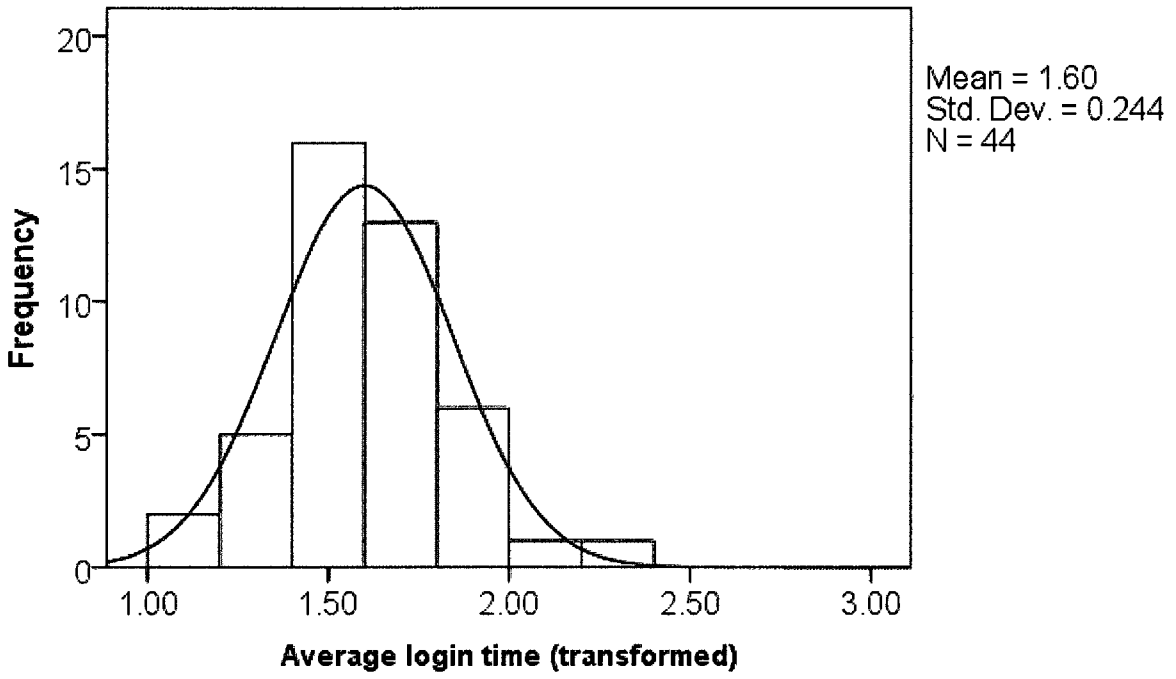


Figure 15. **Frequency distribution (histogram) of log transformed average login times during the second lab session, including all conditions.**

As seen in Figure 15, the transformation rendered the distribution closer to normality, with more acceptable skewness and kurtosis values (skewness = .68, kurtosis = 1.20). The mean of the transformed distribution ($M = 1.60$, $SD = 0.24$) includes data from the object images condition ($M = 1.45$, $SD = 0.19$), the face images condition ($M = 1.60$, $SD = 0.13$), and the house images condition ($M = 1.86$, $SD = 0.22$). Since the transformed data more adequately satisfied the assumptions of ANOVA, the test parametric test was then applied to the data.

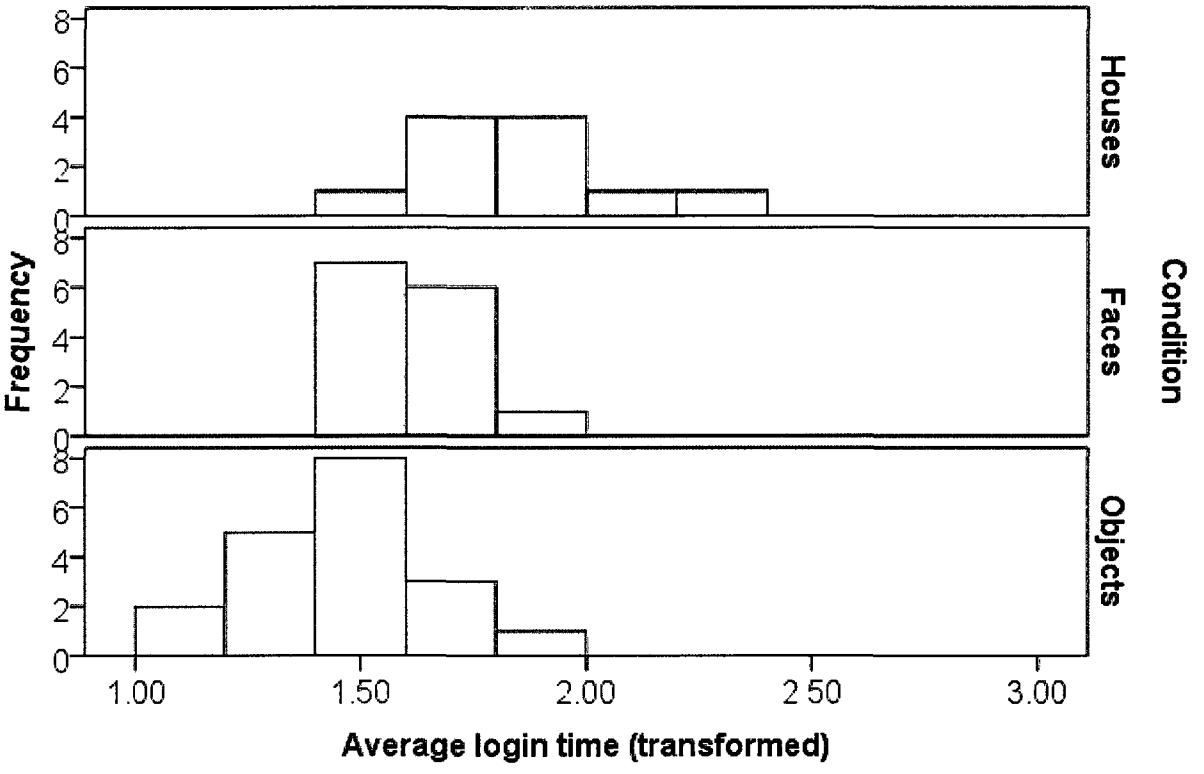


Figure 16. **Frequency distributions (histograms) of log transformed average login times across conditions during the second lab session.**

The analysis of variance on the transformed data (seen in Figure 16) produced significant differences between groups, $F(2,41) = 17.63$, $p < .001$. In order to determine more precisely where those differences were, we then moved to post hoc analyses. The Tukey HSD test (see Table 7) indicated that the transformed mean login time for those in house images condition was significantly greater than that for either of the other two conditions. However, while the transformed mean login time for those in the object images condition was less than that for those in the face images condition, the difference fell short of significance ($p = .065$).

Table 7. Post hoc multiple comparisons between conditions for transformed average login times during the second lab session.

Multiple Comparisons

Tukey HSD

(I) condition	(J) condition	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Houses	Faces	.26303*	.07387	.003	.0834	.4427
	Objects	.41242*	.06946	.000	.2435	.5813
Faces	Houses	-.26303*	.07387	.003	-.4427	-.0834
	Objects	.14939	.06458	.065	-.0076	.3064
Objects	Houses	-.41242*	.06946	.000	-.5813	-.2435
	Faces	-.14939	.06458	.065	-.3064	.0076

*. The mean difference is significant at the 0.05 level.

Recall that the above login time results are from the second lab session. In fact, participants in the object images condition consistently (in every time period) logged in the fastest, followed by those in face images condition, with those in the house images condition typically taking the longest. In order to dispel any doubt that the results for time taken to log in during the second lab session were representative of the results from the other time periods, an analysis of average login times for each condition across all time periods was also conducted. That analysis began by looking at the descriptive statistics and frequency distributions of average time taken to log in across all time periods.

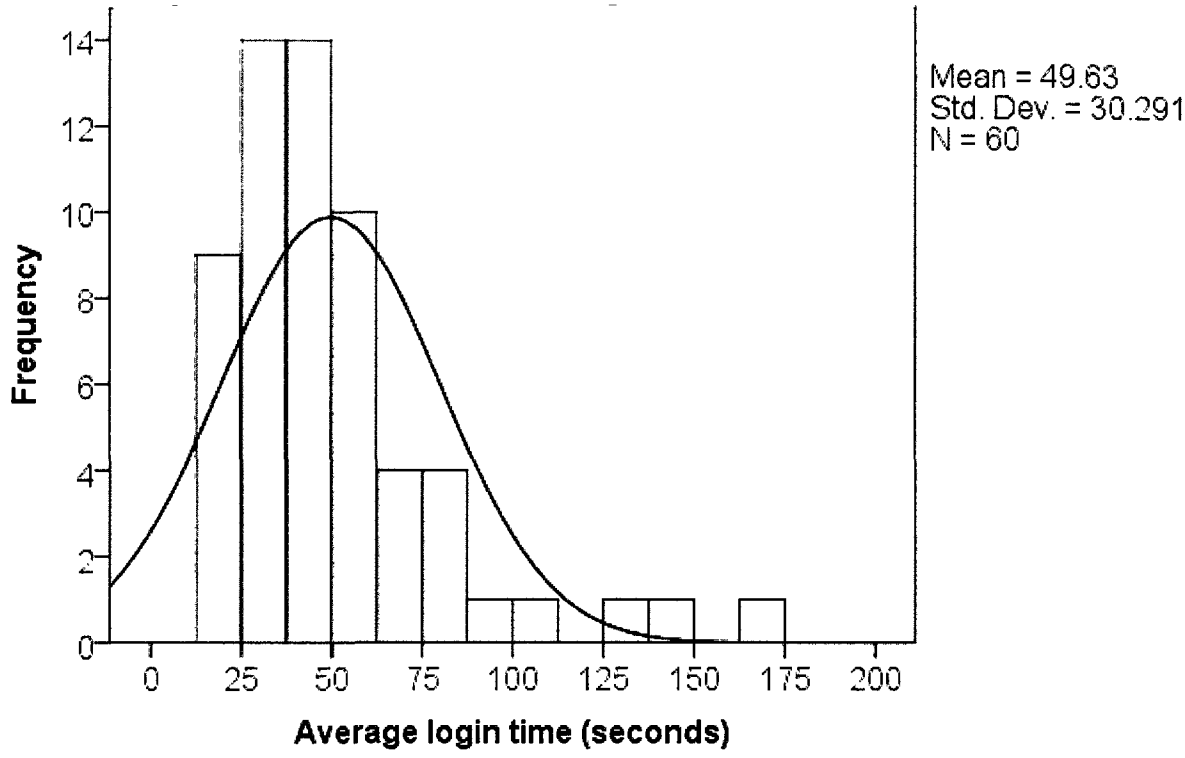


Figure 17. **Frequency distribution (histogram) of average login time across all time periods, including all conditions.**

The frequency distribution represented in Figure 17 includes data from all time periods and all conditions. Descriptive statistics suggest that the assumption of normality has been violated as both the skewness and kurtosis values lie outside the acceptable range (skewness = 2.11, kurtosis = 5.54). Therefore, a Kruskal-Wallis test was used to test the hypothesis. The results of the test were indeed significant $\chi_2(2, N = 60) = 20.95, p < .001$, suggesting a significant difference in time taken to login between conditions.

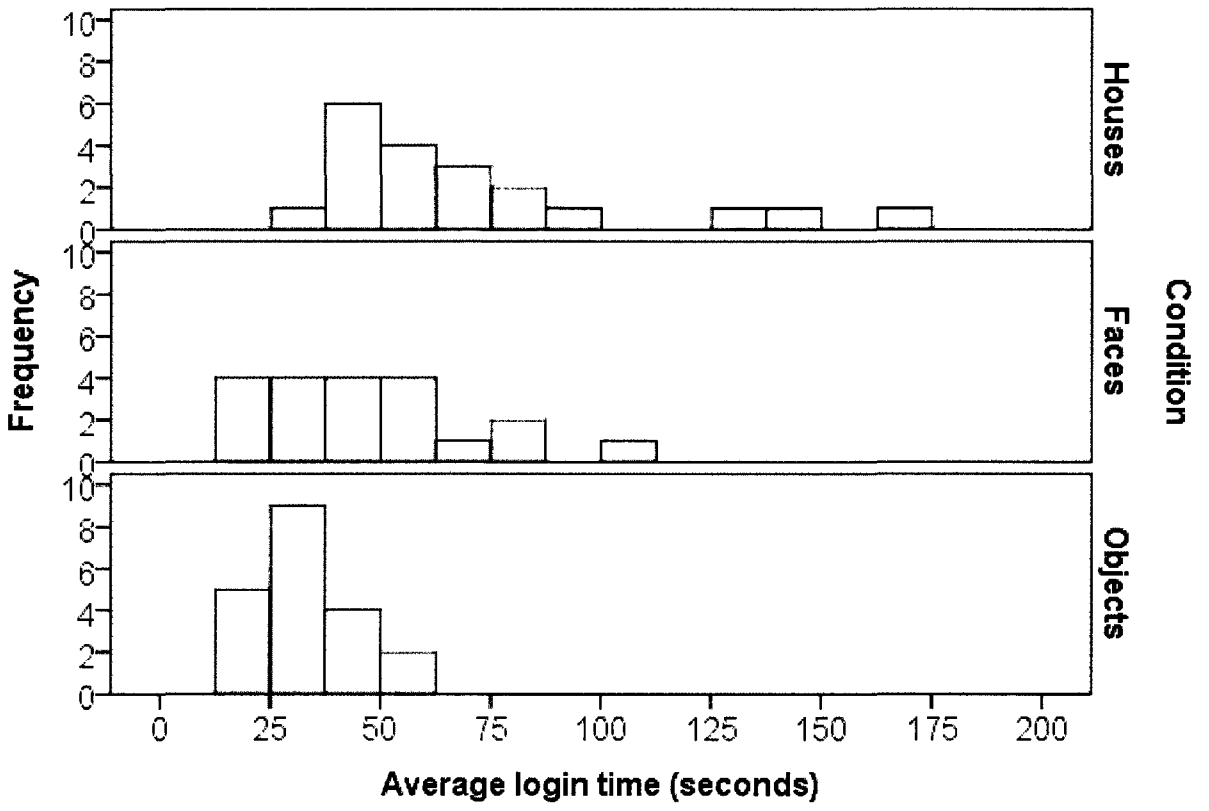


Figure 18. **Frequency distributions (histograms) of average login times across conditions, including all time periods.**

To uncover where the (as seen in Figure 18) significant difference(s) in average login time occurred when including data from all time periods, multiple comparisons were then conducted. Again, the non-parametric test to use when there are two levels in the independent variable is the Mann-Whitney *U* test, with the alpha level corrected by the Bonferroni method ($\alpha = .05 / 3 = .017$). In applying this test to each pair of levels in the independent variable, it was found that while those in the object images condition ($M = 31.77$, $Mdn = 29.83$, $SD = 10.26$) logged in considerably faster than those in the face images condition ($M = 46.60$, $Mdn = 47.05$, $SD = 21.70$), the difference fell short of significance, $z = -2.25$, $p = .025$. Also, while those in the face images condition logged in considerably faster than those in the house images condition ($M = 70.53$, $Mdn = 57.92$, $SD = 38.31$), the difference fell short of significance, $z = -2.39$, $p = .017$.

The only statistically significant difference was between those in the object images condition and those in the house images condition, $z = -4.52$, $p < .001$.Next, the base-10 log transformation was again performed to normalize the distribution. This resulted in a distribution (see Figure 19) that was a better fit for the assumption of normality necessary to proceed with ANOVA (skewness = .43, kurtosis = .18).

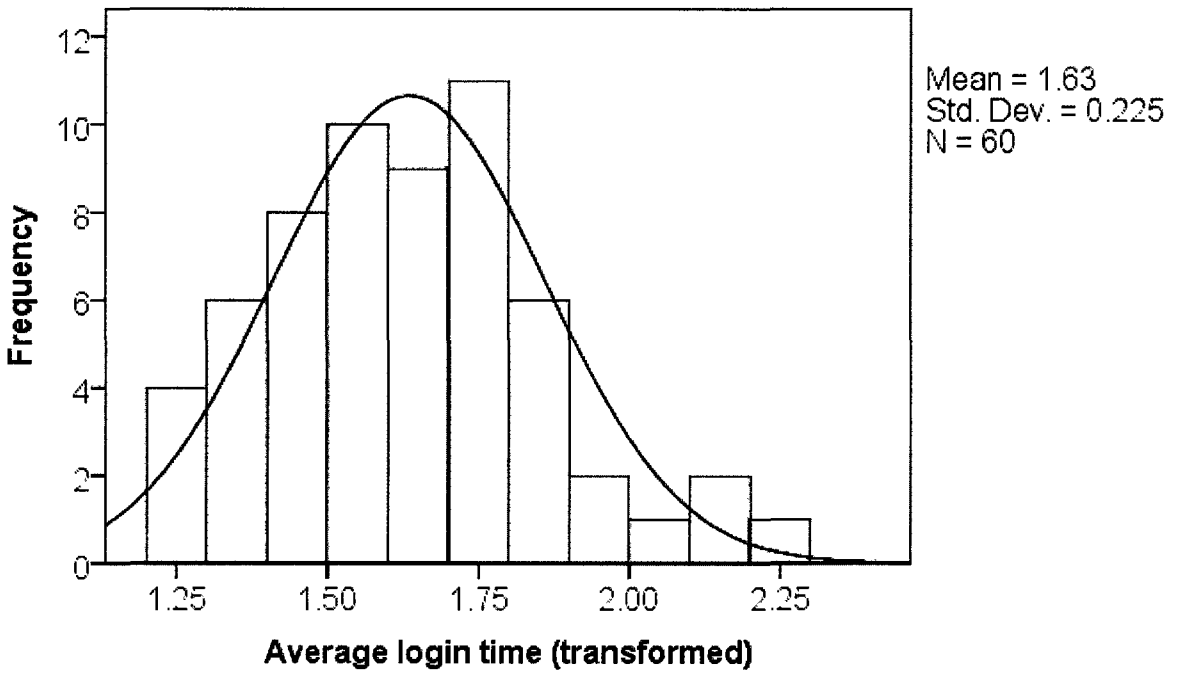


Figure 19. **Frequency distribution (histogram) of log transformed average login times including all conditions and all time periods.**

The mean of the transformed distribution ($M = 1.63$, $SD = 0.23$) included data from the house images condition ($M = 1.80$, $SD = 0.20$), the face images condition ($M = 1.62$, $SD = 0.21$), and the object images condition ($M = 1.48$, $SD = 0.13$). Since the transformed data more adequately satisfied the assumptions of ANOVA, we then proceeded with the parametric test.

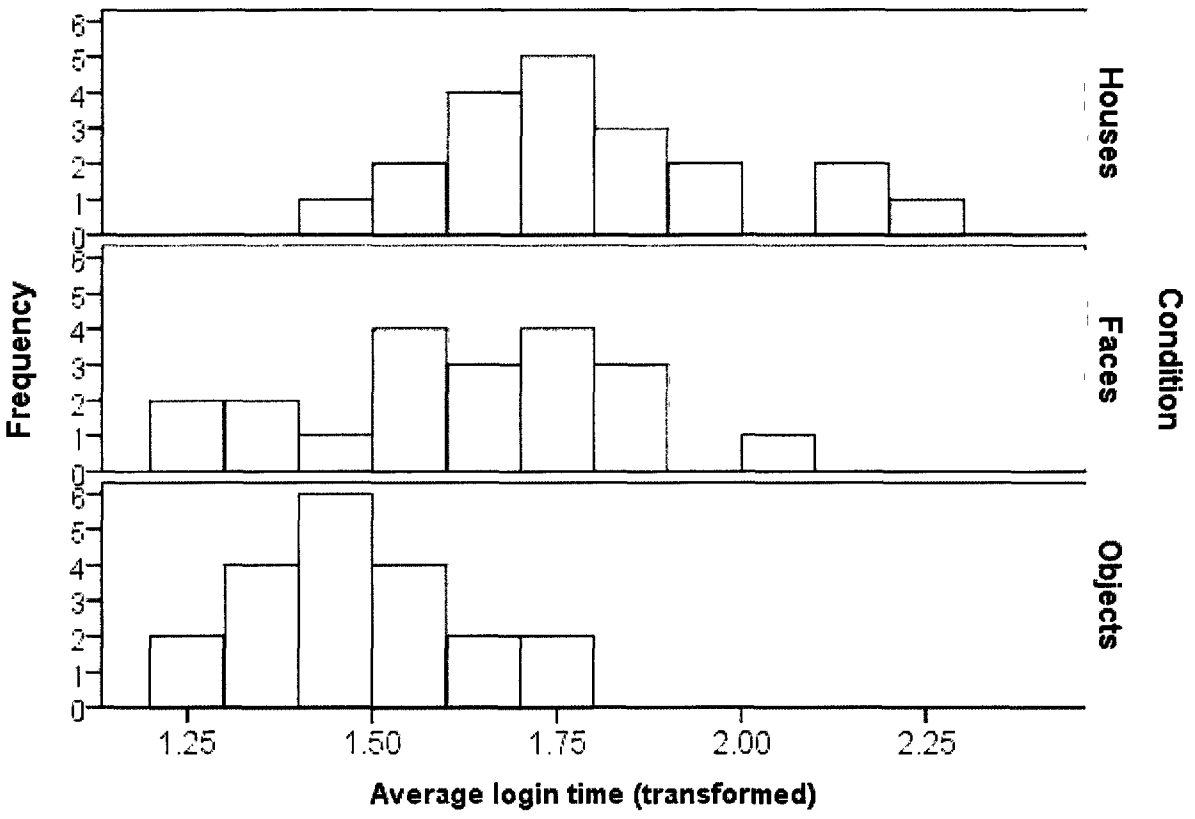


Figure 20. **Frequency distributions (histograms) of log transformed average login times across conditions, including all time periods.**

The analysis of variance on the transformed data (as seen in Figure 20) produced significant differences between conditions, $F(2,58) = 14.78$, $p < .001$. In order to determine more precisely between which conditions those differences were, we then moved to post hoc multiple comparisons. The comparison of transformed average login time (including data from all time periods) between those in the object images condition and those in the face images condition saw differences fall just outside the threshold for significance ($p = .051$). The Tukey HSD test indicated significant differences between all other levels of the independent variable (see Table 8, below).

Table 8. Post hoc multiple comparisons between conditions for transformed average login times, including all time periods.

Multiple Comparisons

Tukey HSD

(I) condition	(J) condition	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Houses	Faces	.17753*	.05863	.010	.0364	.3186
	Objects	.31804*	.05863	.000	.1769	.4591
Faces	Houses	-.17753*	.05863	.010	-.3186	-.0364
	Objects	.14051	.05863	.051	-.0006	.2816
Objects	Houses	-.31804*	.05863	.000	-.4591	-.1769
	Faces	-.14051	.05863	.051	-.2816	.0006

*. The mean difference is significant at the 0.05 level.

In summary, significant differences between conditions in average login time were found both during the second lab session and when including data from all time periods. Significant differences were found by non-parametric tests and by parametric tests on transformed data. Participants in the object images condition logged in the fastest, followed by those in the face images condition, with those in the house images condition taking the longest to log in.

Lastly, we looked for an effect of gender on average login time. There was no significant effect of gender on average login time during the second lab session (z = -.40 , p = .69), or when including data from all time periods (z = -.21 , p = .84).

Questionnaire Feedback

Two questionnaires (see Appendices C and D) were given to participants during the study. The first questionnaire was filled out after the first lab session and the second questionnaire was filled out after the second lab session. Questionnaire items featured ten-point Likert scales ranging from "Strongly Disagree" (0) to "Strongly Agree" (10). To interpret the

feedback, questions were grouped into themes, and the scores for some questions were reversed to create consistency in each theme. Next, each participant's mean rating for the questions in each theme was calculated. As each participant belonged to one of the conditions in the study, the participant means were then used to calculate mean theme ratings for each condition.

The first theme asked participants to compare the graphical passwords used in the study to more traditional text passwords. Questions 3, 8, 12, 14 and 20 from the first questionnaire, along with questions 2 and 3 from the second questionnaire were grouped into this theme. As seen in Table 9, participants generally preferred text passwords over the graphical passwords they used in the study. Note the slight differences between conditions — those in the house images condition preferred text passwords the most, followed by those in the object images condition, with those in the face images having the weakest preference for text passwords.

Table 9. Mean ratings across conditions for the first questionnaire theme. High score suggests: I prefer text passwords.

Theme 1	Mean	SD
Houses	8.26	1.17
Faces	6.86	1.28
Objects	7.27	1.41

The second theme of questions asked participants whether or not they thought that logging in with their passwords took too long. Questions 1 and 19 from the first questionnaire were grouped into this theme. The results in Table 10 show a rather neutral score, suggesting that participants (even those in the house images condition) were unsure whether or not logging in took too long. While the mean ratings for the face and object image conditions were just below neutral, the mean rating for the house images condition was notably higher.

Table 10. Mean ratings across conditions for the second questionnaire theme. High score suggests: Logging in takes too long.

Theme 2	Mean	SD
Houses	5.20	1.85
Faces	4.30	1.60
Objects	4.40	1.54

The third theme asked participants whether or not they felt that the security of the graphical password system that they used was adequate. Questions 2, 10, 11, and 15 from the first questionnaire asked participants how guessable they thought such passwords would be, and whether or not they felt safe using them. As seen in Table 11, all participants were moderately satisfied with the security of the scheme. Those in the house images condition felt notably more secure — likely a result of the increased difficulty they experienced trying to use their *own* passwords.

Table 11. Mean ratings across conditions for the third theme. High score suggests: I trust the security of the passwords.

Theme 3	Mean	SD
Houses	7.24	1.66
Faces	6.65	1.40
Objects	6.51	1.78

The fourth theme of questions had to do with the usability of the scheme. Questions 4 and 17 from the first questionnaire simply asked participants whether it was easy or hard to log in with their passwords. Ratings were again rather neutral (see Table 12). However, participants did

find that it was hardest to log in with house images, followed by face images, with object images being the easiest to log in with.

Table 12. Mean ratings across conditions for the fourth questionnaire theme. High score suggests: The passwords are hard to use.

Theme 4	Mean	SD
Houses	5.85	1.65
Faces	5.05	1.56
Objects	4.05	1.75

The fifth theme among questionnaire items included questions 5 and 18 from the first questionnaire, along with question 4 from the second questionnaire. Participants were asked whether or not they felt that the type of images in their graphical password helped them remember it. As seen in Table 13, participants in the house images condition did not think that house images were helpful. Ratings for the other two conditions were more neutral, with participants in the object images condition liking their image type the most.

Table 13. Mean ratings across conditions for the fifth questionnaire theme. High score suggests: Image type in the passwords was helpful.

Theme 5	Mean	SD
Houses	3.65	1.46
Faces	5.43	1.54
Objects	5.77	1.86

The sixth and last theme among questionnaire items included questions 6 and 21 from the first questionnaire along with question 5 from the second questionnaire, and asked participants how willing they would be to use such an authentication system in their real lives. While the ratings from participants in the house images condition suggested that they would rather not use

such a system, ratings in the other conditions were more neutral (see Table 14). Curiously, participants in the face images condition were most willing to use the graphical password scheme in real life.

Table 14. Mean ratings across conditions for the sixth questionnaire theme. High score suggests: I would use the passwords in real life.

Theme 6	Mean	SD
Houses	3.37	1.83
Faces	5.58	1.83
Objects	4.78	2.39

In summary, participants in the house images condition were not fond of the system they used. Participant feedback for the other two conditions is more interesting. While participants in the object images condition thought that their scheme was slightly more usable, it was the participants in the face images condition that were slightly more willing to use scheme in their everyday lives.

Discussion

The authentication scheme tested in the study had usability *and* security goals. For the purposes of the present study, the number of login errors made by participants, the memorability of the passwords and overall login success of participants, as well as time taken to log in were all considered measures of usability. With regard to security, recall that our design involved a vast increase over traditional cognometric authentication schemes. However, the security of the scheme was only raised high enough to reach equivalence with the least demanding levels of security acceptable for text passwords. While this level of "password strength" is marginally acceptable, questions about usability are evident in the data and feedback from our participants. Interestingly, the only possible exception was the condition in which participants used images of

objects in order to authenticate. Participants in the object images condition made fewer errors, used fewer password resets, were more successful in remembering their passwords, and took substantially less time to log in. Each of these findings will be discussed in further detail, however it may be that the combination of these is what lead participants filling out the post-test questionnaire to rate the object images condition as the easiest to use.

Our first hypothesis test explored the difference between conditions when it came to the amount of login errors committed by participants. Recall that the dependent variable for in this test was the average amount of failed login attempts leading up to either a login success, or the participant giving up. The first important finding from this hypothesis test is that participants in the object images condition made the fewest login errors. However, statistical analysis revealed that the only significant difference in login errors committed during the second lab session was between the object and house image conditions. While results did not show a consistently significant superiority of the object images condition over the face images condition, they certainly provided no evidence that face images are the best type to use in such an authentication scheme.

The misleading nature of the results for average amount of login errors committed by those in the house image condition must be addressed. Participants in the house images condition had a very difficult time in our study, relative to those in the other two conditions. The results from the second lab session show they made the most login errors. However, the between conditions comparisons in prior time periods tell a different story. In those comparisons, participants in the house images condition appear to make much fewer errors when attempting to log in, and are even comparable with those in the objects condition. Interestingly, the reason for the difference has to do with the number of times the participants in the house images condition

tried to log in. During the second lab session, participants complied with requests to at least *try* to login with each of their passwords. However, during the two prior time periods participants were attempting to log in from home, or from the location of their choosing (outside the lab). Participant compliance with our e-mail requests to log in remotely was low for all conditions in our study, but particularly so for those in the house images condition. Not only did they make fewer login attempts, but they were much quicker to "give up" after one failed attempt than those in the other conditions. It seemed that those in the other conditions were more confident that they could eventually get their password right if they kept trying. Of course, this heavily affected the average number of login errors made by those in the house images condition when trying to authenticate from outside the lab. They made fewer errors because they were less willing to try. In the lab they felt more pressure to try, and because of the difficulty of the condition, they made more errors. For this reason, the login errors results for those in the house images condition coming from the second lab session are considered to be a much more accurate depiction than measurements taken during the week.

Certainly, the difficulty experienced by participants in the house images condition has extremely negative implications for its usability. The house images condition was difficult enough to create such doubt in its participants that they did not even want to try. This was most evident when, during the second lab session, participants in the house images condition would make one or two attempts at one password, and then turn to the experimenter and admit that they had forgotten all their passwords. In these instances, which occurred numerous times, the experimenter would say something like, "Don't feel bad, but we ask that all participants at least make an attempt to login with each of their passwords." Wincing, participants would make their best guess. Particular care was given to participants in the house images condition during the

debriefing. They were shown the other conditions in the study, were told that most participants in the house images condition had difficulty, and were reminded that it was the usability of the authentication scheme that was being tested, not the ability of individual participants. It is worth mentioning that when shown each image type during the debriefing at the end of the study, most participants expressed preference for the object images condition. This tendency suggests that there would have been advantages to a within-subjects design, where each participant would have at least one password with each image type.

Participation rates in time periods involving website access from outside the lab were low for all conditions. However, one can also note a consistent decline in participation from the object images condition to the face images condition, with those in the house images condition participating the least from outside the lab. This suggests that not only did the participants who did participate "give up" faster in the house images condition (thus appearing to make fewer errors), they also seemed less willing to participate at all. It is worth noting how the pattern of this *missing data* between conditions adds to the consensus from other findings in the study.

The second hypothesis test looked for differences between conditions in the login success of participants. The data for participant login attempts, successes, and failures was complicated and somewhat sporadic. As a result, there was a need for a simple summary statistic that would provide a straight answer to the most important question – did participants remember their passwords? Again, it was decided that the simplest solution was to assign participants with a "login success score" of 0-3 in each time period, representing the number of passwords they remembered in that time period. By comparing the distributions of these login success scores between conditions, it was possible to analyze differences in login success between conditions.

There was a significant difference between conditions in login success one to two days after the first lab session, and during the second lab session occurring at the end of the study. Importantly, the same trend occurred in every time period. Participants in the object images condition consistently had the highest login success scores, followed by those in the face images condition, with those in the house images condition having the least success. Although pair-wise comparisons revealed that the only significant difference in login success was between those in the object images condition and those in the house images condition, there is something to be said for the consistency in the general direction of the findings.

Again, participants in the object images condition made fewer login errors, had higher login success scores, remembered their passwords longer (on average), had to reset their passwords less often, and as is about to be discussed, took less time to log in. Furthermore, this was the case in every time period. If differences between the two conditions among the greater population were truly insignificant, it would be less likely that one condition would "outperform" the other in every measure used in the study. These findings certainly do not provide any evidence that face images are the ideal type to use in such systems. Lastly, the small sample sizes and large variance values in the present study may have obscured the true relationship between conditions in the study. However, the overall consensus in the data suggests that with a larger sample, differences between conditions may become more pronounced as measures become more accurate.

The third hypothesis test explored differences between conditions in how long it took participants to enter their passwords. This was an important test, as users generally do not want to spend large amounts of time in the authentication process. If graphical passwords are to be seriously considered as a usable alternative to text passwords, they will have to satisfy the user's

expectation for what a reasonable amount of time spent in the authentication process is. Post-test questionnaire feedback showed that participants in every condition felt that it took a bit too long to authenticate with the recognition based password scheme. Although time taken to successfully log in is more of a usability measure than any measure of memory, the more time participants spent looking for the images in their password, the greater the opportunity for distractor interference would become. When participants took longer to log in, it was partly because they had to look through more images before locating and recognizing those from their password set. Participants reported that this interference from familiar images (which they had previously focused on) that were not in their password had a disruptive effect on their ability to recognize the correct images.

The test for average time taken to log in found statistically significant differences both during the second lab session and when aggregating login times from all time periods in the study. Generally speaking, participants in the objects condition always logged in the fastest, typically followed by those in the face images condition, with those in the house images condition typically taking the most time to log in. However, both the parametric and non-parametric analyses used in the hypothesis test told the same story — that the only consistently significant difference in login time was between the object and house image conditions.

There were two important findings from the third hypothesis test. First, the directionality of average login time across conditions had the same implications for the differential usability of the conditions as other findings in the study. Furthermore, in at least one instance (during the second lab session), participants in the object images condition logged in (on average) *significantly* faster than participants in either of the other two conditions. Although participants in the object images condition outperformed those in the other two conditions in every measure

of the analysis, this was the only case in which the differences between the object and face image conditions was significant. For that reason, the relatively short login times of participants in the object images condition during the second lab session served as the best (indeed, perhaps the *only*) irrefutable piece of evidence suggesting the superiority of that image type for use in such authentication schemes.

The general consensus in the directionality of the results suggests that face images may not be the best type of images to use in cognometric graphical passwords. While participants certainly performed better with face images than they did with house images, any evidence of superiority over the object images condition was completely absent. The special way in which people process and thus recognize the faces of others, along with the fact that we all bear a certain expertise in doing so, is certainly advantageous when using a recognition-based authentication system featuring face images. However, it is possible that different advantages afforded by other image types are better suited to the task — the object images used in the present study *may* serve as an example.

The object image type in the present study has certain advantages for memory and recognition that are less available when looking at images of different faces. For one, the images in the object image condition look vastly different from one another. Each object has its own shape and size. Perhaps most importantly, high quality images of colourful objects that stood out from a uniform white background were chosen for the object images condition. That the colours were so bright and so different for each image may have been of great service to participants in the object images condition. This stands in contrast to the images used in the face images condition, which featured much less variation in shape, size, and colour. The second way that each image in the object images condition is unique has more to do with the vast semantic

difference between images. The object images condition featured images of tools, toys, food, flowers, stationery items, furniture, and more. Although faces of varying age, race, gender, expression, etc. bear some semantic difference to the viewer, these differences are dwarfed by the semantic difference between any two categories of objects in the object images condition, and a given object image password in the present study was likely to feature images from several such categories.

One may speculate that the greater visual and semantic differences between images in the object images condition allowed participants to find and recognize specific images more quickly. More importantly, these greater visual and semantic differences may have resulted in a more *distinct* experience for the user when looking at each image. Gallo, Meadow, Johnson, and Foster (2008) discuss distinctiveness in terms of the complexity and uniqueness of the perceptual features in the stimulus, claiming that "distinctive features potentially help subjects to differentiate picture memories, and also provide more features that can subsequently be recollected" (p. 1097). It is the complexity of the differences between random objects that renders their perceptual features so unique. In other words, "distinctiveness can be influenced by the uniqueness of semantic or conceptual features in memory, holding the perceptual input relatively constant" (p. 1109). The images in the object images condition had greater distinctiveness than the images in the other two conditions, and yet their presentation was the most constant. The uniform presentation of the object images, each one centered and isolated on a white background, is not to be underestimated. Similar to how Hunt (2006) describes distinctiveness as difference in the context of similarity (see page 18), Gallo *et al.* (2008) stipulate that certain perceptual constants are upheld. The uniform, almost exhibit-like

presentation of the objects (along with their categorization *as* objects) creates this constant context in which the vast perceptual and semantic differences between objects are more obvious.

Once again of importance is Craik and Lockhart's (1972) levels-of-processing framework, and more specifically how it relates to distinctiveness. Recall that, according to the levels-of-processing framework, the method and depth of processing affects how an experience is stored in memory. Gallo *et al.* (2008) argue that "the levels of processing effect on memory is based on recollective distinctiveness" (p. 1109), whereby the *distinctiveness* of an experience is what determines our ability to remember it. Furthermore, that "uniqueness or distinctiveness within a level of processing can influence performance (is) consistent with many other studies highlighting the importance of distinctiveness for recall and recognition" (p. 1109). Gallo *et al.,* emphasize that it is the uniqueness of the stimulus during encoding and retrieval that plays such an important role in memory. The authors argue that "the most effective way to encode items for subsequent recall or recognition is to associate each item with information from pre-existing knowledge that can later provide a large number of unique features to retrieve" (p. 1109). The relevance of this argument is not to be overlooked. When comparing our ability to remember faces of people we do not know to our ability to remember objects that most of us know a great deal about, there is a great difference in the amount of pre-existing knowledge available for association.

In the above description by Bruce and Young (1986) of the different types of information available when looking at a picture of someone's face, *identity-specific* information was said to contain all the information that cannot be derived by simply looking at the picture. It could be argued that the most important aspects of knowledge pertaining to the subject matter of such a picture — a person — is not available when simply looking at it. This is *not* typically the case

when looking at an image of an orange, for example. Most people have had all kinds of first (and second) hand experiences with oranges through which they have learned the information most pertinent to them. That is to say, most people know more about "everyday" objects than they know about people they have never seen or heard of before. As a result, there is a greater opportunity for people to associate images of those objects with their own knowledge and/or experience.

Another important difference between image types had more to do with the strategies reported by participants to help themselves remember their passwords. In the house and face image conditions, participants often reported that they created some brief verbalization of each image in their password. For example, they would come up with names such as "sporty blonde" for a certain face image or "the barn house" for a certain house image. This labeling process seemed less important for those in the object images condition, perhaps because each object already had its own name. This could be said to relate to the previous point about association — that because there are less unknowns to the subject matter in the object images, there is greater opportunity for associations to be made. Regardless, it seemed that the verbalization of images (particularly in the house and face image conditions) was used by participants to assist the *recall* of the images in their password. Often, participants would attempt to recall the verbalizations of the images in their graphical password before any attempt to recognize them on the screen. In instances where recall was unavailable, participants would then resort to relying purely upon recognition as they scanned through each panel of images. It was during this pure recognition task that interference from familiar distractor images (that may have been relevant for another password) seemed to be the strongest.

In summary, there may be certainly usability advantages to using face images in recognition-based graphical passwords. However, other image types may have advantages of their own, and contrary to what is claimed by those at *PassFaces*, the present study found no evidence that images of faces are the most advantageous type to use in the cognometric graphical password systems. Although images of random "everyday" objects were the superior image type in the present study, more research is necessary to determine whether or not there is an ideal image type to use in such schemes.

There are two implications for visual memory in general. First, expertise with the subject matter of the visual stimulus is advantageous for the memorability of that stimulus. Second, the more *distinct* that the experiences resulting from a visual stimulus are, the more distinctly it is processed, which facilitates later memory. Interestingly, it could be argued that these two ideas are related, as the benefits of expertise may be driven by an enhanced ability to make important distinctions among certain types of subject matter. To put it candidly, whether or not houses have faces may depend on your expertise with houses. An architect or real estate agent might be inclined to suggest that indeed they do.

There were several important limitations (and resulting lessons) to the findings in the present study. Although ecological validity is sorely lacking from a great deal of previous laboratory research, and the current study attempted to increase validity by including at-home testing, great care must be taken in the experimental design of ecologically valid studies to anticipate important problems such as non-compliance and missing data. Studies with small sample sizes are particularly vulnerable to such problems, which can lead to statistical limitations in the interpretation of results. The effects of and relationships between variables become obscured by a lack of data, limiting the power and accuracy of findings. However, the

consistency in the general direction of the findings of the present study may substantiate the need for reliability testing with a much larger sample and similar variables.

Only half of the participants in the best condition of the study (the object images condition) remembered their passwords the whole week. Thus, it would seem that the methods used to increase the security of the scheme in the present study rendered it less usable than similar schemes with much lower security (theoretical password space), such as the one (Passfaces) reported on by Brostoff and Sasse (2000). As such, one direction for further research is an exploration into the effects of adjusting the balance between usability and security. Specifically, there may be other ways to increase the security of cognometric authentication schemes without sacrificing usability. For example, the random placement of the important images among distractors in each panel unnecessarily decreases the usability of the scheme. The original purpose of having the "password images" appear in random locations was to protect the user against shoulder surfing attacks. However, this feature only protects against the weaker forms of shoulder surfing. Sophisticated shoulder surfing attacks will involve an actual video recording of the screen during authentication — randomizing image location among each panel would have no effect. Importantly, if image location were *not* randomized, the memorability of the scheme might be much greater. Users would have the opportunity to remember the general location of the important image in each panel, making it much easier to locate and recognize the images in their password. In fact, it is possible that such a change would also afford an increase in panel size, which would increase the security of the scheme. One might also wonder whether differences in image type become less important when the user has this extra positional information about the images in their password.

Related to the above is the question of whether it would be a greater burden on users to increase panel size, or to simply add more panels. The usability of cognometric authentication schemes with smaller theoretical password spaces, like the five 3x3 panels of images used by Everitt *et al.* (2009) and most typically by *PassFaces*, is known to be relatively higher than the usability found of the scheme in the present study. However, as previously explained (see pages 14 - 16) the security of such schemes is unfortunately low. At question is whether increasing panel size or increasing the number of panels (to increase security) has a greater negative effect on usability.

In one of the conditions of a separate study by Chiasson *et al.* (2009) on graphical passwords, participants were tasked with remembering three *text* passwords of their choosing for up to six days. That study used the same websites and study framework as the present study. The text passwords were required to have at least one digit and one upper-case letter, and be a minimum of six characters long. On average, those participants remembered their text passwords for approximately 104 hours. The time for them to enter correct passwords after six days was approximately six seconds. Importantly, the password memorability findings in the study on text passwords are comparable to those of the present study. Specifically, participants in the face images condition of the present study also remembered their passwords for 104 hours (on average). Participants in the object images condition remembered their passwords 23% longer, with a mean memory time of 135 hours (a difference of 31 hours). This is corroborated by the comparable proportion of passwords remembered by participants for the duration (approximately one week) of each study. However, even participants in the object images condition (the condition with the fastest logins) of the present study took substantially longer ($M = 31.03$) to log in than those in the text passwords condition of the study by Chiasson *et al.* ($M = 6.40$). Recall

that the findings from the second theme in the questionnaire feedback suggest that participants in all conditions of the present study felt that logging in took too long.

Conclusion

Graphical passwords are a promising alternative to text passwords as they have the potential to solve traditional usability issues without sacrificing security. Cognometric graphical passwords are a specific type in which users must remember a set of images and later recognize them as they appear among grids of distractor images. The research question in the present study was whether or not there is an effect of image type in such schemes. To address this question, the present study assigned cognometric graphical passwords with an acceptable security level and three different image types to participants, and monitored their ability to use them over the period of one week. The results of the study indicate that image type was found to have a significant effect on the amount of login errors, the login success, and the login times of participants. Furthermore, the specific level of theoretical password space (security) of the graphical passwords in the present study showed similar memorability to a comparable text password study, but much greater (and potentially problematic) password entry time.

The findings were limited, however, by a number of factors, including the relative novelty of graphical passwords compared to the many years of practice most people have had with text passwords, that participants were not truly invested in the success of their login attempts, that a small sample size was further reduced by a between-subjects design and missing data. However, it is possible that with training, practice, and the right properties (image type, image location, panel size, number of panels, etc.), cognometric graphical passwords could achieve a balance between usability and security that is superior to that of traditional text

passwords. However, if these authentication schemes are to be considered as a viable alternative to traditional text passwords, further work is necessary in order to determine their optimal configuration.

References

Ashby, G., Isen, A., & Turken, A. (1999). A neuropsychological theory of positive affect and its influence on cognition. *Psychological Review*, *106*(3), 529-550.

Bousfield, W. A., Esterson, J., & Whitmarsh, G. A. (1957). The effects of concomitant colored and uncolored pictorial representations on the learning of stimulus words. *Journal of Applied Psychology*, *41*, 165-168.

Bower, G. H. (1972). Mental imagery and associative learning. In L. Gregg (Ed.), *Cognition in Learning and Memory* (51-88). New York: Wiley

Bower, G. H., Karlin, M. B., & Dueck, A. (1975). Comprehension and Memory for Pictures. *Memory and Cognition*, *2*, 216-220.

Brostoff, S., & Sasse, M. (2000). Are Passfaces more usable than passwords? A field trial investigation. *British Human-Computer Interaction Conference (HCI)*, September 2000.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305-327.

Calkins, M. W. (1898) Short studies in Memory and Association from the Wellesley College Laboratory. *Psychological Review*, *5*, 451-462.

Chiasson, S., Biddle, R., & Van Oorschot, P. C. (2007). A second look at the usability of click-based graphical passwords, *3rd Symposium on Usable Privacy and Security (SOUPS)*, July 2007.

Chiasson, S., Forget, A., Biddle, R., & Van Oorschot, P. C. (2008). Influencing users toward better passwords: Persuasive Cued Click-Points. *Human Computer Interaction (HCI), the British Computer Society*, September 2008.

Chiasson, S., Forget, A., Stobert, E., Biddle, R., & Van Oorschot, P. C. (2009). Multiple password interference in text and click-based graphical passwords. *ACM Computer and Communications Security (CCS)*, Chicago, USA, November 2009.

Chiasson, S., Deschamps, C., Stobert, E., Hlywa, M., Freitas Machado, B., Chan, G., & Biddle, R. (2009). *The MVP Web-based Authentication Framework*, Technical Report TR-10-19, School of Computer Science, Carleton University, Ottawa, Canada.

Chiroro, P. M., Tredoux, C. G., Radaelli, S., & Meissner, C. A. (2008). Recognizing faces across continents: The effect of within-race variations on the own-race bias in face recognition. *Psychonomic Bulletin & Review*, *15*(6), 1089-1092.

Craik, F., & Lockhart, R.S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior*, *11*(6), 671-84.

Cranor, L. F., & Garfinkel, S. (2005). *Security and Usability: Designing Secure Systems That People Can Use*. Sebastopol, CA: O'Reilly.

Davis, D., Monrose, F., & Reiter, M. K. (2004). On User Choice in Graphical Password Schemes. *Proceedings of the 13th USENIX Security Symposium*, 151-164.

De Angeli, A., Coventry, L., Johnson, G., & Renaud, K. (2005). Is a picture really worth a thousand words? Exploring the feasibility of graphical authentication systems. *International Journal of Human-Computer Studies*, *63*(1-2), 128-152.

Dhamija, R., & Perrig, A. (2000). Déjà vu: A user study using Images for Authentication. *Proceedings of the 9th Conference on USENIX Security Symposium*, 4-4.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: an effect of expertise, *Journal of Experimental Psychology: General*, *115*, 107–117.

Dirik, A. E., Memon, N., & Birget, J. 2007. Modeling user choice in the PassPoints graphical password scheme. In *Proceedings of the 3rd Symposium on Usable Privacy and Security* (Pittsburgh, Pennsylvania, July 18 - 20, 2007). SOUPS '07, vol. 229. ACM, New York, NY, 20-28.

Everitt, K. M., Bragin, T., Fogarty, J., & Kohno, T. (2009). A comprehensive study of frequency, interference, and training of multiple graphical passwords. *Proceedings of the 27th international conference on Human Factors in computing systems* (held April 4 – 9, 2009, in Boston, MA, USA). CHI '09. ACM, New York, NY, 889-898.

Mike, M. (2010) the Face of Tomorrow Project. Available online at http://www.faceoftomorrow.com/

Farah, M.J. (1996) Is face recognition "special"? Evidence from neuropsychology. *Behavioural Brain Research*, *76*(1-2), 181-189.

Farah, M.J., Wilson, K. D., Drain, M., & Tanaka, J. (1998). What is "special" about face perception? *Psychological Review*, 105(3), 482-498.

Feldmeier, D., & Karn, P. (1990). UNIX Password Security – Ten Years Later. *Advances in Cryptology – CRYPTO '89* (Lecture Notes in Computer Science 435).

Flechais, I., Riegelsberger, J., & Sasse, M.A. (2004). Divide and conquer: the role of trust and assurance in the design of socio-technical systems. *Technical Report*, 2004.

Fried, I., & Evers, J. (2006, Feb. 14). Gates: End to passwords in sight. RSA Conference on information security. *CNET News*. Retrieved in December 2009, from http://news.cnet.com/2100-7355-6039177.html

Gallo, D. A., Meadow, N. G., Johnson, E. L., & Foster, K. T. (2008). Deep levels of processing elecit a distinctiveness heuristic: Evidence from the criterial recollection task. *Journal of Memory and Language*, *58*, 1095-1111.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Science*, *4*, 223-233.

Hunt, R. R. (2006). The Concept of Distinctiveness in Memory Research. In R. Hunt & J. Worthen (Eds.), *Distinctiveness and Memory* (3-26). New York: Oxford University Press.

Jermyn, I., Mayer, A., Monrose, F., Reiter, M. K., & Rubin, A. D. (1999). The design and analysis of graphical passwords. *Proceedings of the 8th conference on USENIX Security Symposium*, p.1-1, August 23-26, 1999, Washington, D.C.

Johnson, M., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, *40*(1-2), 1-19.

Kausler, D. H. (1974). *Psychology of Verbal Learning and Memory*. New York: Academic Press

Kirita, T., & Endo, M. (1995). Happy face advantage in recognizing facial expressions. *Acta Psychologica*, *89*(2), 149-163.

Klein, D. (1990). Foiling the cracker: A Survey of, and Improvements to, Password Security. *Proceedings of the 2nd USENIX security workshop*, 5-14.

Madigan, S. (1983) Picture Memory. In Imagery, Memory, and Cognition (Hillsdale, NJ: Erlbaum, 1983).

Mandylion Research Labs (2010). *Password Cost Estimator*. Retrieved in November 2009, from http://www.mandylionlabs.com/PRCCalc/PRCCalc.htm

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3-

35.

Minnebusch, D.A., Suchan, B., Koster, O., & Daum, I. (2009). A bilateral occipitotemporal network mediates face perception. *Behavioural Brain Research, 198*(1), 179-185.

Monrose, F., & Reiter, M.K. (2005). Graphical Passwords. In L. F. Cranor, & S. Garfinkel (Eds.), *Security and Usability* (pp. 157-174). Sebastopol, CA: O'Reilly Media, Inc.

Morris, R., & Thompson, K. (1979). Password security: a case history. *Communications of the ACM, 22*(11), 594-597.

National Institute of Standards and Technology NIST (January, 2010). The Color FERET Database. Retrieved from http://face.nist.gov/colorferet/

Nelson, D. L., Reed, U. S., & Walling, J. R. (1977) Pictorial Superiority Effect. *Journal of Experimental Psychology: Human Learning and Memory, 3*, 485-497.

Paivio, A., Rogers, T. B., & Smythe, P. C. (1968) Why Are Pictures Easier to Recall Than Words? *Psychonomic Science, 11*, 137-138.

Paivio, A., & Csapo, K. (1969). Concrete image and verbal memory codes. *Journal of Experimental Psychology, 80*, 279-285.

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review, 76*, 241-263.

Paivio, A. (1973). Picture superiority in free recall: Imagery or dual coding? *Cognitive Psychology. 5*(2), 176-206.

Passfaces Corporation, "The science behind Passfaces," White paper, http://www.passfaces.com/enterprise/resources/white_papers.htm, accessed November 2009.

Renaud, K. (2009). On user involvement in production of images used in visual authentication. *Journal of Visual Languages and Computing*, *20*(1), 1-15.

Roediger, H. L. (2008). Relativity of Remembering: Why the Laws of Memory Vanished. *Annual Review of Psychology*, *59*, 225-254.

Saltzer, J., & Schroeder, M. (1975). The protection of information in computer systems. *Proceedings of the IEEE*, *63*(9), 1278-1308.

Sasse, M. A., & Flechais, I. (2005). Usable Security. Why do we need it? How do we get it? In Cranor, & S. Garfinkel (Eds.), *Security and Usability* (pp. 13-30). Sebastopol, CA: O'Reilly Media, Inc.

Shepard, R. N. (1967). Recognition Memory for Words, Sentences, and Pictures. *Journal of Verbal Learnings and Verbal Behavior*, *6*, 156-163.

Smith, S. (1987). Authenticating Users by Word Association. *Proceedings of the 31st Annual Meeting of the Human Factors Society*, 135-138.

Sporer, S. L. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law*, 7(1), 36-97.

Standing, L. (1973). Learning 10,000 Pictures. *Quarterly Journal of Experimental Psychology*, *25*, 207-222.

Stock.xchang (2010) The leading free stock photography website. Available online at http://sxc.hu/

Thorpe, J., & Van Oorschot, P. C. (2007). Human seeded attacks and exploiting hot-spots in graphical passwords, in *16th USENIX Security Symposium*, August 2007.

Valentine, T. (1999). And evaluation of the Passface personal authentication system. Goldsmiths College University of London, Tech. Rep., February 1999.

Van Oorschot, P. C., & Thorpe, J. (2008). On predicting and exploiting hot-spots in click-based graphical passwords. *School of Computer Science, Carleton University, Tech. Rep.* TR-08-21, November 2008.

Watkins, M. J. (1974). When is recall spectacularly higher than recognition? *Journal of Experimental Psychology*, *102*(1), 161-163.

Weirich, D., & Sasse, M.A. (2001). Pretty good persuasion: a first step towards effective password security for the real world. *Proceedings of the New Security Paradigms Workshop 2001* (sept. 10-13, Cloudcroft, NM); (ACM Press 2001), 137-143.

Wiedenbeck, S., Waters, J., Birget, J.C., 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.

Wu, T. (1999) A Real-World Analysis of Kerberos Password Security. *Proceedings of the 1999 ISOC Symposium on Network and Distributed System Security.*

Yan, J., Blackwell, A., Anderson, R., & Grant, A. (2005). The Memorability and Security of Passwords. In L. F. Cranor & S. Garfinkel (Eds.), *Security and Usability: designing secure systems that people can use* (pp. 129-142). Sebastopol, CA: O'Reilly.

Zhao, W., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (2003). Face Recognition: A Literature Survey. *ACM Computing Surveys*, *35*(4), 399-458.

Appendix A: Consent Form – Graphical Passwords Usability Test

Research Personnel

Max Hlywa Principle Investigator Carleton University (613) 520-2600 Ext. 6317 mhlywa@connect.carleton.ca	Dr. Robert Biddle Faculty Sponsor Carleton University (613) 520-2600 Ext. 6317 robert_biddle@carleton.ca

Purpose

The purpose of this usability test is to evaluate the usability, security, and effectiveness of graphical passwords. We are also trying to determine what types of images are ideal for use in graphical passwords.

Task Requirements

You will be asked to complete a set of tasks that include using graphical passwords and filling out a short questionnaire. In the coming week, you will also be asked (via e-mail) to use graphical passwords to access a website from the location of your choice. At the end of the week, we will briefly meet again and you will be asked to fill out a second questionnaire.

Duration and Locale

Both the first and second sessions should take less than one hour. Between sessions, you will access a website from the location of your choice. Upon completion of the second session you will receive a $20 honorarium OR course credit for your time. Sessions will take place at the HotSoft lab located in room HCI 2110.

Potential Risk/Discomfort

There will be no psychological or physical risk in this study.

Anonymity/Confidentiality

All data that is collected will be held strictly confidential. The data will only be made available to those people involved with this testing, and will be coded for identification purposes to maintain anonymity.

Right to Withdraw

You have the right to withdraw at any time, without any explanation as to the reason for withdrawing from the testing. You will receive the $15 honorarium or course credit even if you choose to withdraw from the study.

If you have concerns about the ethics of this research, please contact Dr. Monique Sénéchal. For other questions about the research, please contact Dr. Janet Mantler:

Dr. Janet Mantler Chair, Department of Psychology Carleton University (613) 520 2600 ext 4173 psychchair@carleton.ca	Dr. Monique Sénéchal Chair, Carleton University Ethics Committee for Psychological Research Carleton University (613) 520 2600 ext 1155 monique_senechal@carleton.ca

Signatures

I have read and understand the above terms of testing and I understand the conditions of my participation. My signature indicates that I agree to participate in this experiment.

Participant's Name: _____

Participant's E-mail: _____

Participant's Signature: _____

Researcher's Name: _____

Researcher's Signature: _____

Date: _____

Appendix B: Participant Information – Usability Test - Student

This information will be held completely confidential. **(Please do not put your name on this form!)**
Age: _____ years
Sex: (check one) ☐ male ☐ female

At what level are you studying?
☐ Undergraduate ☐ Masters ☐ Ph.D ☐ Other

What year of study are you in? _____

In what academic program are you enrolled?

On a scale of 1 (novice) to 10 (expert), how would you rate yourself with respect to your computer skills?

Novice									Expert
1	2	3	4	5	6	7	8	9	10

How often do you browse the web?

☐ Daily ☐ Several times a week ☐ Once a week ☐ Less than once a week

How many passwords do you have? (Please answer with a number)

On a scale of 1 (never) to 10 (always), how often do you re-use passwords that you also use elsewhere?

Never									Always
1	2	3	4	5	6	7	8	9	10

What criteria do you use for choosing a password? (Select more than one if appropriate)

☐ It is easy for you to remember
☐ It is suggested by the system
☐ It is difficult for others to guess
☐ It is the same as another password you currently have
☐ It is related to the system or website

How concerned are you about the security of your passwords?

Not at all concerned									Very concerned
1	2	3	4	5	6	7	8	9	10

If you had to create a new password for your bank account (using a normal text password system) because you had forgotten or misplaced your old password, how would you go about choosing your new password?

Have you ever used a graphical password (using pictures to enter a password instead of typing in letters and numbers)? If so, where?

Participant Information – Usability Test – Non-Student

This information will be held completely confidential. **(Please do not put your name on this form!)**

Age: _____ years

Sex: (check one) ☐ male ☐ female

What is your occupation?

On a scale of 1 (novice) to 10 (expert), how would you rate yourself with respect to your computer skills?

Novice									Expert
1	2	3	4	5	6	7	8	9	10

How often do you browse the web?

☐ Daily ☐ Several times a week ☐ Once a week ☐ Less than once a week

Approximately how many web sites do you visit that require a username and password? (Please answer with a number)

On a scale of 1 (never) to 10 (always), how often do you re-use passwords that you also use elsewhere?

Never									Always
1	2	3	4	5	6	7	8	9	10

What criteria do you use for choosing a password? (Select more than one if appropriate)

☐ It is easy for you to remember
☐ It is suggested by the system
☐ It is difficult for others to guess
☐ It is the same as another password you currently have
☐ It is related to the system or website

How concerned are you about the security of your passwords?

Not at all concerned									Very concerned
1	2	3	4	5	6	7	8	9	10

If you had to create a new password for your bank account (using a normal text password system) because you had forgotten or misplaced your old password, how would you go about choosing your new password?

Have you ever used a graphical password (using pictures to enter a password instead of typing in letters and numbers)? If so, where?

Appendix C: Graphical Passwords Post-Task (Session 1) Questionnaire

1. With practice, I could quickly enter graphical passwords

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

2. I would trust a graphical password to protect my financial information.

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

3. Graphical passwords are quicker to use than text-based passwords

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

4. Logging on using a graphical password was easy

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

5. It was easy to distinguish images in my password set from the other images

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

6. I would use a graphical password

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

7. If I didn't log in to my account for a few weeks, I would still remember my graphical password

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

8. It would be easier to remember 5 different text passwords than 5 different graphical passwords

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

9. Given the choice between a text password and a graphical password, I would choose a graphical password

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

10. My accounts would be secure if protected by a graphical password

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

11. Graphical passwords would be easy for attackers to guess

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

12. Logging on using a graphical password was easier than with a text password

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

13. Text passwords are more secure than graphical passwords

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

14. I prefer text passwords to graphical passwords

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

15. It would be easy to guess my graphical password

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

16. Graphical passwords are easy to remember

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

17. It was difficult to enter my password

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

18. The type of images in my graphical password made them easier to remember

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

19. Graphical passwords are too time-consuming

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

20. If I was in a hurry, I would rather enter a text-based password than a graphical password

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

21. I would be happy if computer systems used graphical passwords instead of text passwords

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

What strategy did you use for remembering your graphical passwords?

Which images were the easiest to remember? Why?

Did your strategy change because you had several graphical passwords to remember? If so, how?

If you could choose what type of images to use, what would you choose? Why?

Appendix D: Graphical Passwords Post-Task (Session 2) Questionnaire

1. Graphical passwords are easy to remember

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

2. Graphical passwords are easier to remember than text passwords

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

3. I would prefer to log in using text passwords

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

4. The type of images in my graphical password helped me remember it

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

5. I would like to be able to use graphical passwords in my real life

Strongly Disagree									Strongly Agree
1	2	3	4	5	6	7	8	9	10

Did you use any strategies to help you remember your graphical passwords? If so, what strategies did you use?

Do you usually use any strategies to remember your text passwords? If so, what strategies do you use?

Which images were most helpful in remembering your passwords? Why?

Appendix E: Debriefing Form – Graphical Password Usability Test

The research we are conducting is part of a larger study examining the usability, practicality, and security of graphical passwords. These usability studies aim to assess the effectiveness of graphical passwords as alternatives to text-based passwords for systems requiring user authentication.

In this study, we are comparing the usability of graphical passwords to that of text passwords. We are also interested in how the type of the image used in the graphical password effects memorability. For example, in one of the conditions of the study we are using images of faces. We are interested in whether facial recognition offers any advantage over the recognition of other images. One hypothesis is that certain types of images, particularly those of faces, are easier to remember than other types.

The results of the usability test will be used to make recommendations on how graphical passwords can be improved. Your thoughts, comments, and opinions will be taken into consideration in making design recommendations. We would like to thank you for participating in this usability test. Your time and effort are greatly appreciated.

If you have any further questions regarding this research, please contact:

Max Hlywa Principle Investigator Carleton University (613) 520-2600 Ext. 6317 mhlywa@connect.carleton.ca	Dr. Robert Biddle Faculty Sponsor Carleton University (613) 520-2600 Ext. 6317 robert_biddle@carleton.ca

If you have concerns about the ethics of this research, please contact:

Dr. Janet Mantler Chair, Department of Psychology Carleton University (613) 520 2600 ext 2664 psychchair@carleton.ca	Dr. Monique Sénéchal Chair, Carleton University Ethics Committee for Psychological Research Carleton University (613) 520 2600 ext 1155 monique_senechal@carleton.ca

Please keep this form until next time so that you remember your username and next appointment time.

Username: _____

Appointment: _____

Appendix F: Reminder e-mail to participants

Dear participant,

As was discussed in the first session of the graphical password study that you participated in, we request that you access and use our secure website, by clicking the links below.

http://vip.soft.carleton.ca/bestof/
http://vip.soft.carleton.ca/univ101/
http://vip.soft.carleton.ca/photos/

We would also like to remind you of your upcoming appointment for the second session of the graphical password study, after which you will be rewarded for participation.

You have a booked appointment for: _____

If you can no longer attend this appointment or would like to change times, please contact us at:

mhlywa@connect.carleton.ca

Thank you very much, and we appreciate your participation.

Max Hlywa
Human Oriented Technology Lab
Carleton University

Appendix G: Recruitment Poster

EARN $20 !!!

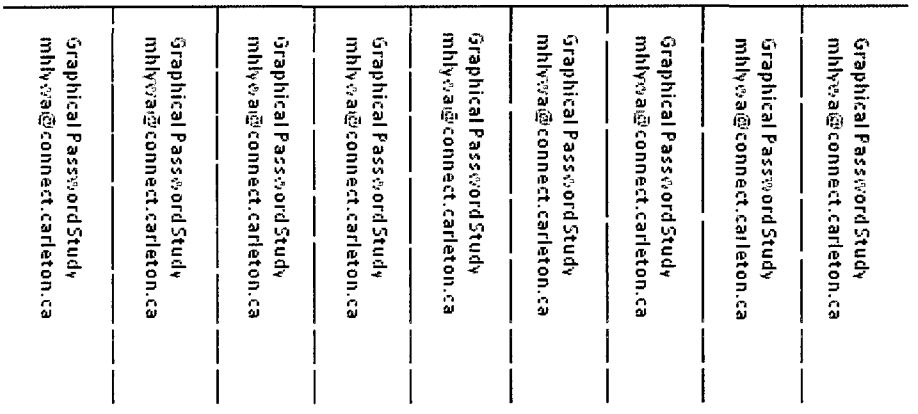
or course credit

Graphical Password Study

- must be able to use keyboard and mouse
- 2 part study: 2 sessions one week apart
- both sessions less than one hour

To sign up, or for more info. contact:

mhlywa@connect.carleton.ca



Appendix H: Image Examples

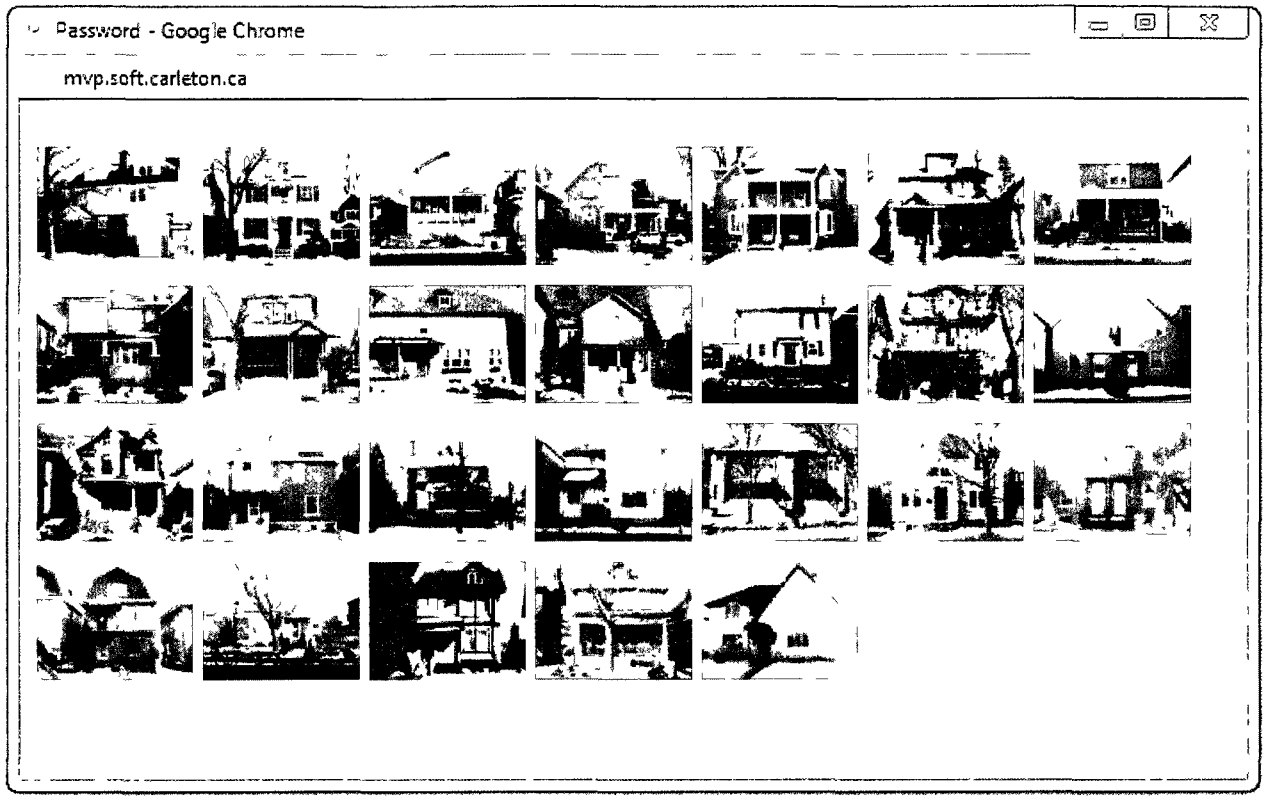


Figure 21. **Example images from the house images condition.**

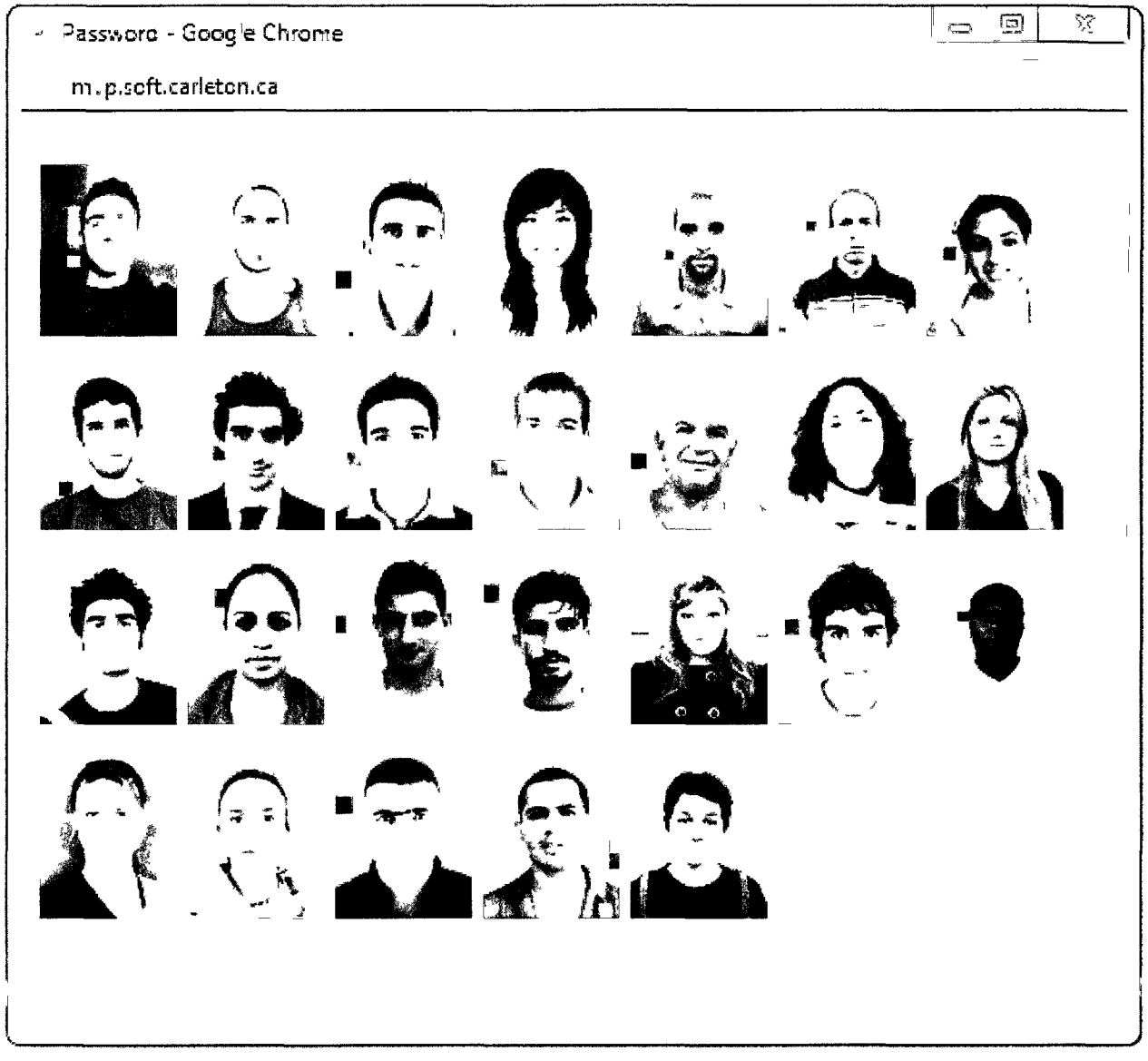


Figure 22. **Example images from the face images condition.**

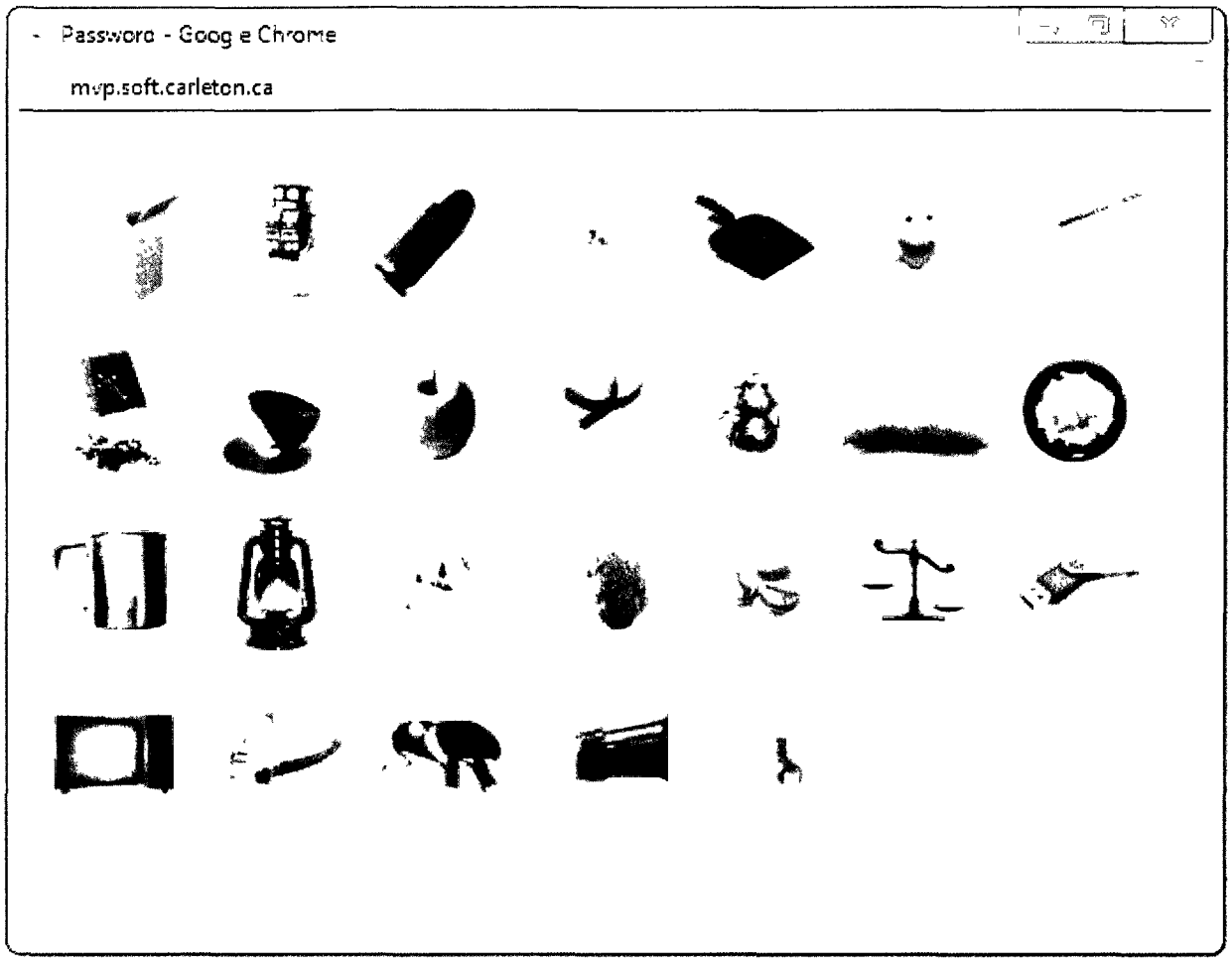


Figure 23. **Example images from the object images condition.**

Appendix I: Website Examples

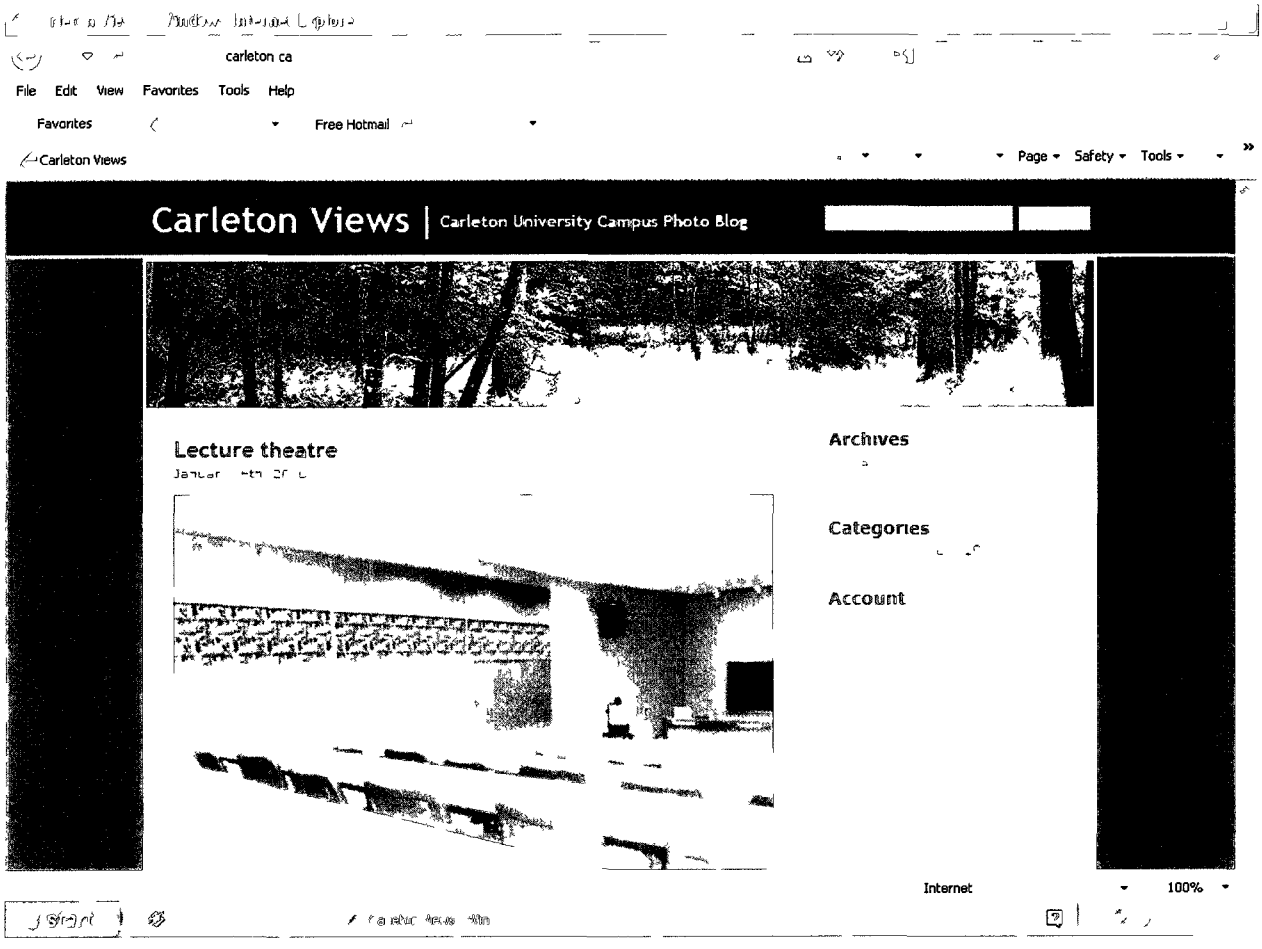


Figure 24. **Screengrab from the photo blog website.**

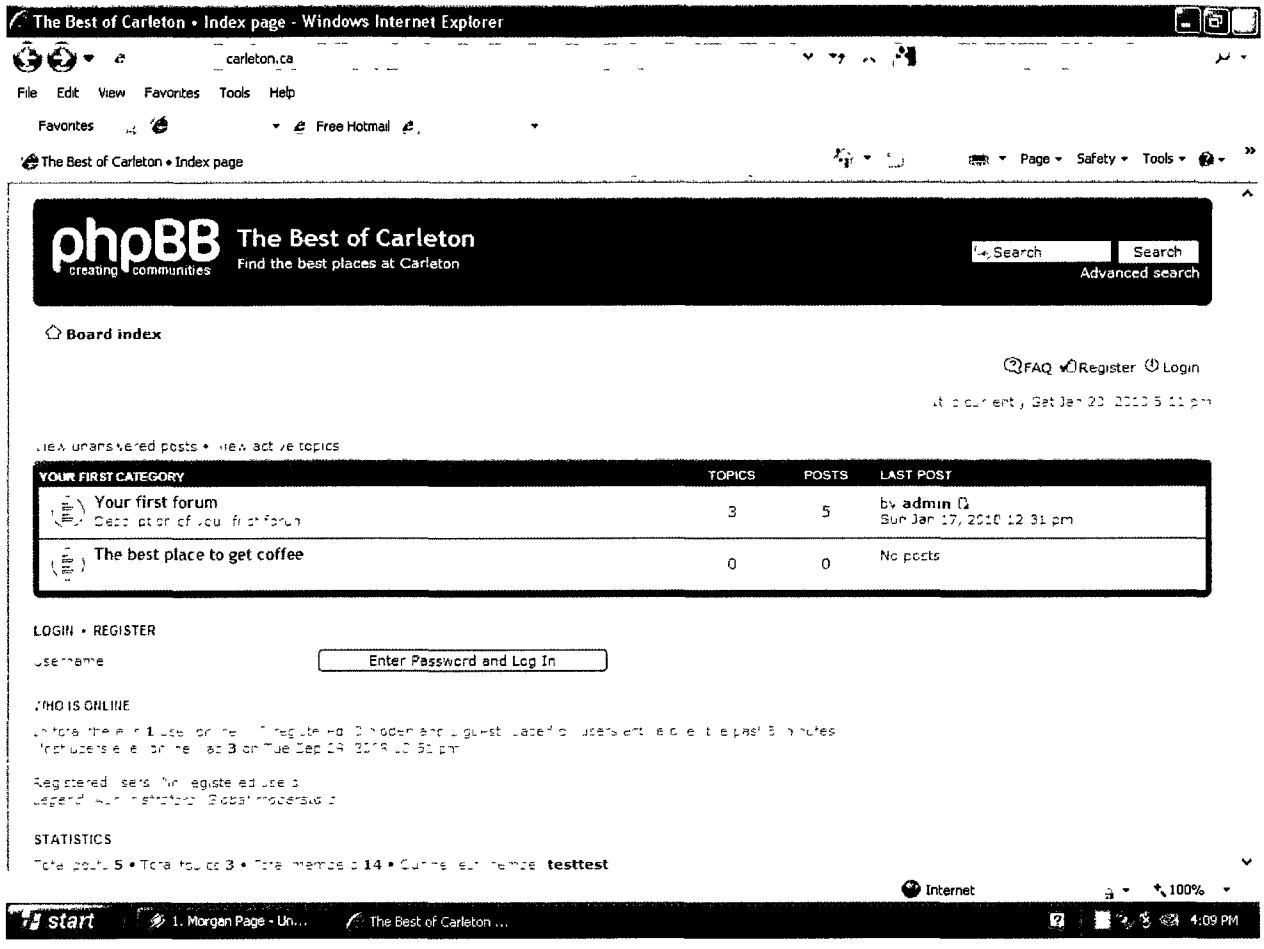


Figure 25. **Screengrab from the message board website.**

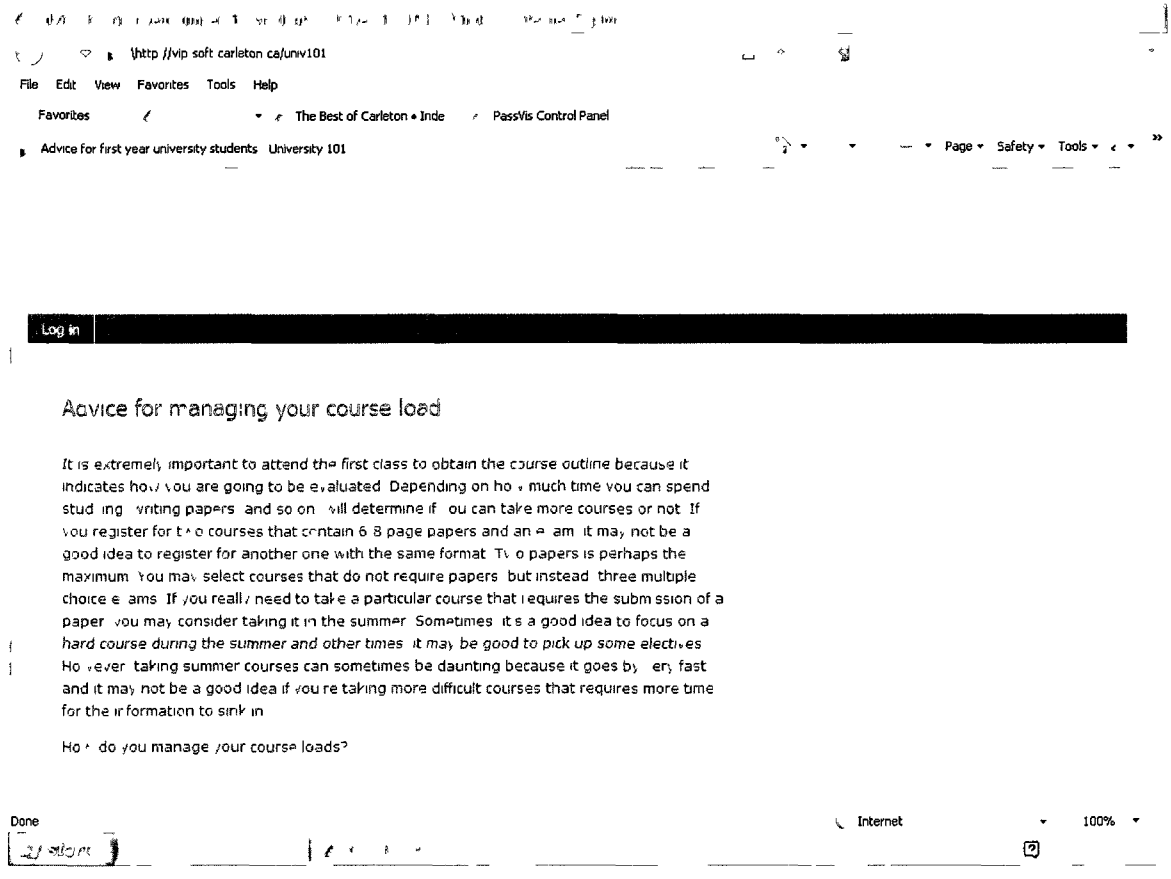


Figure 26. **Screengrab from the blog website.**