DO YOU SEE YOUR PASSWORD?
APPLYING RECOGNITION TO TEXTUAL PASSWORDS

by

Nicholas Wright

A thesis submitted to
the Faculty of Graduate and Postdoctoral Affairs
in Partial Fulfillment of the requirement for the degree

Master of Arts

in

Psychology

Carleton University
Ottawa, Canada

Your file Votre référence
ISBN: 978-0-494-83120-5
Our file Notre référence
ISBN: 978-0-494-83120-5

Abstract

Text-based password systems are the authentication mechanism most commonly used on computer systems. Graphical passwords have recently been proposed because the pictorial-superiority effect suggests that people have better memory for images. The most widely advocated graphical password systems (which use images) are also based on recognition rather than recall. This approach is favored because recognition is a more effective manner of retrieval than recall, exhibiting greater accuracy and longevity of material. However, schemes such as these combine both the use of graphical images and the use of recognition as a retrieval mechanism. This thesis seeks to address this confound by exploring the recognition of text as a novel means of authentication. We hypothesized that there would be significant differences between text recognition and text recall conditions. The results show that the conditions are comparable, and not significantly different in memorability, but text recognition required more time to authenticate successfully.

Acknowledgements

The completion of this work would not have been possible without the support of several people, to whom I owe many thanks. First I would like to thank my supervisor Dr. Robert Biddle, whose thoughtful guidance and wisdom was truly inspiring. My co-supervisor, Dr. Andrew Scott Patrick's careful critiques of my work helped assure a thorough and comprehensive review of the theory and a concise explanation of the experiments and results. I would also like to thank my cohorts in HotSoft and the HOTLab for their support and encouragement over the course of my studies, and my friends and family members for their love and patience throughout this endeavor.

Table of Contents

List of Tables

List of Figures

List of Appendices

Introduction

Authentication, in the context of computer security, is the practice of identifying oneself in order to acquire access to information or resources. In today's computing environment, the vast majority of user authentication is accomplished using text password mechanisms (Angeli, Coventry, Johnson & Renaud, 2005). In text password systems, the valid user is required to submit a secret password, which only they should know, in order to verify their identity to a computing system. Ideal passwords would be those that are easy for users to remember, simplifying the process of authentication, but difficult for attackers to guess, rendering the system secure (Yan et al., 2005).

In striving for ideal passwords, we are introduced to the security / usability tradeoff. Strong (secure) passwords are difficult to remember, and passwords that are easy to remember are typically weak. Systems requiring passwords that are too strong result in frequently forgotten, reset and disclosed passwords, and these systems inspire password reuse and the creation of passwords that are minimally secure. Weak passwords, while easy to remember, are vulnerable to a wide variety of attacks ranging from shoulder surfing to brute force and dictionary attacks, as we describe in the sections below. The intent of this study is to explore potential improvements to text passwords for authentication.

Many of today's current practices related to text password systems are based on assertions made by credible sources, and industry best practices have evolved based on them (Burr, Dodson, Polk & Evans, 2004; Florencio & Herley, 2010). Unfortunately these best practices have evolved with little regard to human factors research, focusing instead on security. For example, password-based systems often insist that new passwords be composed of at least eight characters, use both upper and lowercase letters, at least one number and one special

character, that they be changed frequently, and that they not resemble previously used passwords (FIPS, 1985). Rules such as these are often too difficult for users to observe without disclosing their passwords. In fact, several authorities in the area of computer security have suggested that users are "the weakest link" in the computer security chain (Ferguson & Schneier, 2000). Human factors practitioners do not refute the data, but rather take a more positive stance and suggest that security mechanisms be designed to account for human characteristics in an effort to enhance security (Sasse, Brostoff & Weirich, 2001). Human factors research may be able to inform designs that are sensitive to the limitations of human cognitive ability, while simultaneously acting to increase the level of security.

Passwords are a form of knowledge-based authentication, where access depends on "something you know". There are other authentication mechanisms available for use in situations requiring secure authentication (van Oorschot, Vanstone & Menezes, 1996). Systems can validate users based upon possession of physical tokens such as RFID tags or banking cards, which are dependent on "something you have". Alternatively, other systems can authenticate based upon scanning a person's body, which is referred to as biometric measurement. Examples of biometric authentication include fingerprints or iris scans, and these are "something you are." The benefits of password-based authentication mechanisms include a low cost to implement and maintain because they require only commonly available hardware, and passwords are easy to initialize and reset. Passwords are also very portable, since there is no extra token to carry from place to place, and passwords avoid the privacy-related concerns associated with the use of biometric data (Herley, van Oorschot & Patrick, 2009).

Text-based password systems do have many weaknesses. People have reported experiencing difficulty recalling them from memory, especially when their composition includes numbers and special characters. The number of passwords a person is required to use regularly,

and frequent password changes, are additional issues for users. In many cases users simply cannot remember their passwords and resort to disclosing them either by writing them down or sharing them with others (Sasse, Brostoff & Weirich, 2001).

A large body of work intended to improve the use of knowledge-based authentication systems involves "graphical passwords" (Suo & Zhu, 2005; Davis, Monrose & Reiter, 2004). These are systems where the "password" or secret is not a word at all, but rather a picture, set of pictures, or picture features. Graphical authentication mechanisms are potentially even more secure than alphanumeric text based password systems, while capitalizing on humans' enhanced memory for images.

The proposals for graphical password systems, however, introduce design features that go beyond the use of pictures instead of text. For example, some systems involve cued recall and others involve recognition. By contrast, text-based systems involve pure recall alone, with the user entering text in a traditional blank input box. Previous studies have shown support for the usability of graphical password systems by comparing cued-recall and recognition-based graphical password systems with the traditional recall-based text password systems (Brostoff & Sasse, 2000). These comparisons involve a confound, however, because the proposed authentication mechanisms benefit from both the pictorial superiority effect and recognition-based retrieval. Text password systems could potentially benefit from the work that has gone into the development of the new recognition-based methods, without resorting to the use of graphical materials. The differences between pure recall (also referred to as free recall), cued recall and recognition processes are described in the Related Theory section.

Graphical Password Systems

A great deal of research in the area of usable security has focused on the design and implementation of graphical password systems. Surveys of the proposed schemes have been provided by Suo & Zhu (2005), Davis, Monrose & Reiter (2004) and Biddle, Chiasson, & Van Oorschot (in press). To date, all implementations of graphical password systems can be categorized as a drawmetric, locimetric or cognometric mechanism (Renaud & De Angeli, 2009).

Drawmetric systems, such as the Draw-a-Secret (Jermyn, Mayer, Monrose, Reiter, & Rubin, 1999) system shown in Figure 1, require users to compose an initial drawing during the set-up phase, which must then be redrawn later in order to authenticate with a given system. This is therefore a pure recall system, similar in nature to text passwords. Moreover, users may experience difficulty with this graphical password system because their drawing must be reproduced with sufficient accuracy for the mechanism to recognize it as being correct and granting the associated permissions. Further analysis of many users' drawings suggested that people tended to compose symmetric drawings as their secret (Thorpe & van Oorschot, 2004). In response, Background Draw-a-Secret systems require that users compose and redraw their secret pattern over a background image (Dunphy & Yan, 2007), rendering it increasingly similar to locimetric mechanisms as described below.

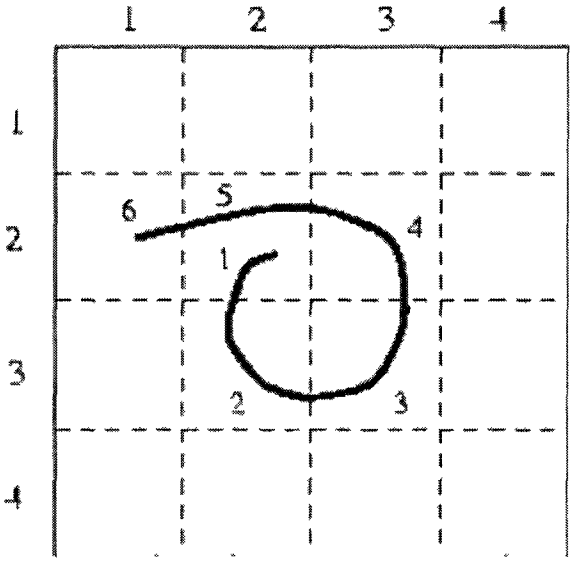


Figure 1: Draw-a-Secret authentication mechanism sample.

Locimetric graphical password schemes are a cued-recall-based method of authentication whereby the system presents users with an image, and locations on the image are selected and recorded for authentication. The initialization phase requests that users select several points on the image, the set of which becomes that user's "password" which must be recalled for future authentication.

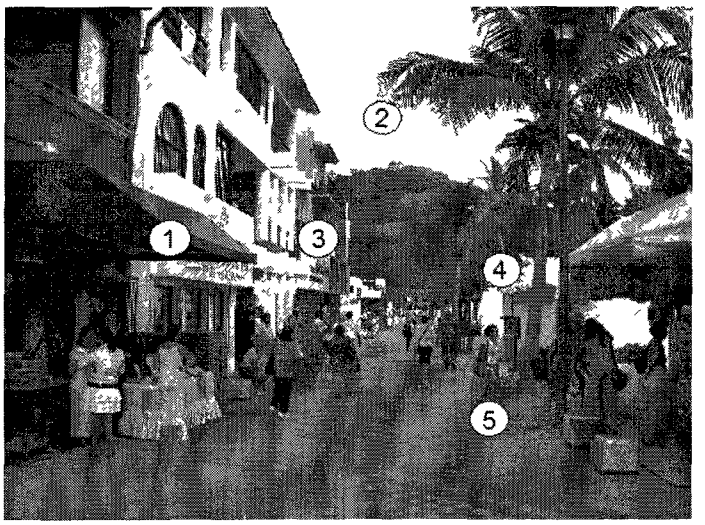


Figure 2: Example locimetric authentication image and authentication points.

Wiedenbeck, Waters, Birget, Brodskiy & Memon (2005) developed the most widely known locimetric graphical password system, *PassPoints*. The usability of this system was established as sufficient for practical deployment (Chiasson, Biddle & van Oorschot, 2007). However, Thorpe & Van Oorschot (2007) discovered that the distribution of chosen click-points for a particular image was far from random, instead centering on "hotspots" – places of increased likelihood of selection. This phenomenon places a significant limitation on the effective password space related to the image, as we discuss in detail in the next section. Persuasive Cued Click-Points (PCCP) was proposed as a modification of the original PassPoints scheme, forcing users to select their click-points from smaller random areas of the image (Chiasson, Forget, Biddle, & van Oorschot, 2008). This adaptation appears to have remedied the hotspot issue while preserving previous login success rates.

Cognometric graphical password systems function by presenting a series of panels of images to the person requesting access, asking them to choose the single correct image on each panel. These recognition-based schemes involve the selection of specific images for password entry. These images are learned upon registration with the system and are plainly presented to the user for recognition at login time. Among existing cognometric schemes, *PassFaces* (PassFaces Corp, 2009) is the most commercialized and studied example. All of the images used in *PassFaces* are of peoples' faces, as depicted in Figure 3 below. Users must select their chosen faces, displayed along with 8 distractor images per panel, over four panels.

Figure 3: Example of a PassFaces login screen.

These systems have been lauded as highly usable (Dhamija & Perrig 2000). User studies have shown that users can remember these types of passwords well, and for long periods of time (Brostoff & Sasse, 2000, Moncur & LePlâtre, 2007). However, the PassFaces example demonstrates relatively weak security, comparable to that of a four-digit bank card PIN (see below for a discussion of password strength).

As was found with locimetric schemes, Davis et al. (2004) identified a weakness regarding the images chosen by users to compose their password in PassFaces. There was a strong tendency to select attractive faces, especially those from one's own race, and this increases the likelihood that an attacker could guess a user's password. The bias inherent in user selection of password images can be guarded against, by assigning users' password images. This is the approach now taken by the commercial PassFaces system. Assigning the faces that

comprise the password raises the possibility of reduced memorability, but no published studies have investigated the associated impact on memory performance.

Cognometric graphical passwords have been suggested as the most promising innovation in knowledge-based authentication because they leverage recognition rather than recall (Renaud, 2010). Since human memory is better able to recognize previously encountered information than it is at recalling material without cues (Kausler, 1974), it is our hope that this principle can be used in the design of improved text-based authentication mechanisms.

Related Theory

Recognition Versus Recall.

Failing to retrieve information from memory is a universal experience. There are three principle mechanisms for accessing information previously acquired. In *recognition*, information is presented to the individual and the individual is asked to make a judgment about whether or not the information is familiar or not. When material is not present to be recognized, it must be recalled from memory. *Recall* can take place with or without the presence of cues. *Cued recall* is a method of retrieving information from memory with the help of a cue, which acts as a hint, aiding the search through memory. Pure, free, or uncued recall is the retrieval of items from memory without any help from the surrounding environment (Crowder, 1976). Providing a cue to an item in memory increases the likelihood of successful recall, and speeds the rate of recall. Uncued recall is the most difficult manner in which known material is retrieved for use (Tulving & Pearlstone 1966).

Ebbinghaus (1885) conducted the first empirical studies of learning and memory, which he conducted upon himself. Based on his research, he discussed the effectiveness of learning, relearning and the ability of human memory, which he had termed "savings". He noted that relearning was 73% more rapid than the initial learning phase. He developed tests (initially conducted upon himself alone) of recall, in which he specified which information was to be recalled, and recognition, in which subjects were asked if they remembered information that may or may not have been presented to them. Lastly, he developed tests of reconstruction in which previously learned material was randomly presented and the task was to organize it according to its original order. These tests were labeled "the four R's" but it was later acknowledged that they are not four distinct ways of testing memory. Reconstruction is equivalent to a recognition task,

with a focus on order, and relearning is closely related to recall, since it is effectively a repeated test of recall.

Luh (1922) and Postman and Rau (1957) conducted widely acknowledged work in this area and confirmed that recognition was the most sensitive measure of memory, boasting higher scores and lesser deterioration in their tests.

Following these findings of differences in performance during the different memory tests, there was a great deal of interest in developing theories that explained the results, and in models of human memory. The first model, called *Tagging* (Yntema and Trask, 1963), stated that items to be remembered were assigned tags related to the semantic meaning and the time of occurrence of the material. In the tagging theory, recognition was merely the verification of an existing tag, whereas recall required an exhaustive search of relevant tags. This theory explained false recognition errors, since items not present on to-be-remembered word lists could be declared recognized due to tagging misfires, and memory degradation was associated with the passage of time since the tag's creation.

The *Strength* model succeeded Tagging (Bernbach, 1967; Wickelgren & Norman, 1966) and proposed that more recently experienced material would have stronger traces in memory than those of less recent material. An analogy for this model was that the link in memory for each word to be remembered was like a bucket of water, where the water level would rise with each presentation, and the level would drop with the passage of time – the higher the water level, the stronger the association for that memory. This model described recognition as a more effective retrieval mechanism than recall because the presentation of the material to be retrieved strengthens the link in memory – raising the water level in the bucket.

The *High-Threshold* theory built upon the Strength model, adding that a critical threshold existed for the memory of each word. If a specific threshold were surpassed then an accurate

decision would confirm the event, and if not, then the answer could be guessed. In this model the threshold for every decision could vary according to the consequences associated with any decision. The theory of signal detectability was in opposition to this, because of the importance of context. It stated that while some arbitrary threshold may exist, there wasn't any critical point associated with a given item, and that familiarity influenced the potential for both recognition and recall (Crowder, 1976).

Hollingworth (1913), and later Tulving (1975), wrote that recognition was accomplished by retrieving the context in which a given item had occurred. Norman (1968) contributed that recall worked in the opposite manner – knowing the context, what was the item? Anderson and Bower (1972, 1974) later added that this was dependent on events that occurred at the time of encoding. Giving emphasis to this, Tulving (1975) developed the theory of Encoding Specificity, which stated that:

> "Specific encoding operations performed on what is perceived determine what is stored, and what is stored determines what retrieval cues are effective in providing access to what is stored." (Tulving & Thomson, 1973)

Tulving's encoding specificity principle stated that retrieval was dependent upon the combination of material stored in memory and certain cues that would facilitate its availability. Cues could be very general, such as "enter your password," or more specific, as is the case with "what is your mother's maiden name?" In either case, successful retrieval relies on the extent to which the cue reinstates the manner in which the information was stored.

There were two early hypotheses on the relation between recognition and recall. First came the threshold-sensitivity hypothesis, which stated that recognition was much like recall, only easier (requiring a lower threshold). Research later discredited this hypothesis because it implied a constant positive correlation between recognition and recall, with recognition always

easier then recall. Variables were identified that produce opposite effects on the performance of these functions, notably word frequency and whether items were intentionally or incidentally learned.

For example, Shepard (1967) showed subjects 540 words, half high frequency, half low. Participants were sequentially shown 60 pairs of words, where one word was part of the previously studied list, and the other not. When presented with each pair, they were asked to decide which word belonged to the previously studied list within each pair. When the previously studied word was high in frequency, correct recognition was around 84%, regardless of the frequency of the distractor. When the known word was low in frequency, correct recognition occurred on average 93% of the time, regardless of the frequency of the distractor. However, Deese (1961) has shown that word frequency produces the opposite effect on one's ability to recall words from memory.

Postman (1964) demonstrated that intentional learning (deliberately learned material) produces better results in recall situations, but Eagle and Leiter (1964; and Estes & DaPolito 1967) showed that incidental learning (material learned incidentally through exposure) produces better results in the case of recognition tests. Weaknesses such as these challenged researchers to identify better models of human memory.

The second main hypothesis, known as the generate-plus-recognize hypothesis, posited that recall is similar to recognition, but with an extra step. Recall was said to require individuals to generate a set of items that could possibly contain the one sought after and then a recognition decision would be made among the items in the set, whereas in recognition the generation phase could be skipped because the item was presented to the individual. According to the Generate-Recognize model, which developed upon this hypothesis (Watkins & Gardiner, 1979), the cue restricts the set of possibilities through which someone would have to search to determine the

answer. This model is strongly aligned with the theory of encoding specificity. It explains the effects on the processes of recognition and recall witnessed in the tests of the variables above by explaining that they can affect the generation or recognition phases independently of each other. If some variable affects recall results positively, but negatively impacts recognition, this can be explained as negatively influencing the recognition sub-process, but positively influencing the generation sub-process.

More recent work (Hintzman, 1990; Richardson-Klavehn & Bjork, 1988; Schacter 1987) has determined differences between declarative, explicit or conscious memory and non-declarative, implicit or unconscious memory. This distinction is strongly supported by studies involving amnesic participants, who have impairments in their abilities to recognize, recall and learn new material, but who are perfectly capable in tests of priming, conditioning and skill learning. Thus it has been acknowledged that memory is not a single system, but consists of multiple elements and processes. Both recognition and recall depend on declarative memory, but recognition also depends on the ability to process the related cues.

The processing of cues can benefit from a phenomenon known as perceptual priming, a process through which detecting and identifying material is facilitated by recent encounters with that same material (Tulving & Schacter, 1990). Thus the ability to recognize words or objects depends not only on a conscious assessment of the material, but also on "increased perceptual fluency" or priming (Gardiner 1988; Jacoby, 1983; Johnston, Dark & Jacoby, 1985). Therefore, recognition capitalizes on encoding in both declarative and non-declarative memory, while recall is limited to declarative memory alone.

Knowledge of these models of memory and retrieval processes leads us to an understanding that recognition and recall situations are handled in different manners by people asked to remember material. The majority of tests performed on memory indicate that our faculty

for recognition consistently produces more effective and persistent results than those of recall. Therefore, it stands to reason that authentication methods capable of taking advantage of our enhanced ability to recognize information are more memorable, and thus more usable, than traditional text passwords relying solely on pure recall.

Password Strength.

There are several manners of attack that motivated parties can use in an attempt to thwart password authentication mechanisms and gain access to restricted information or services. Excluding attackers' manipulation of software vulnerabilities to circumvent the authentication phase altogether, attacks are generally classified as capture or guessing-based attempts to determine actual passwords.

Capture related attacks are those that require interception of the password during entry, or deceiving a user into divulging the secret under false pretense. These include shoulder surfing, reconstruction, malware, phishing, and social engineering methods. Guessing-based attacks involve performing numerous attempts of potentially educated guesses at generating a password to gain access to the protected system. These attacks may be performed systematically, guessing every possible password in what are known as "brute-force" attacks. More refined guessing attacks limit the attempts to "words" found on discrete lists, or "dictionaries". Dictionary attacks can be optimized, for example, guessing more likely words first, and potentially omitting words that are unlikely to be used.

This study does not address capture attacks. Moreover this study does not address ordered dictionary attacks, because the passwords to be used are random and assigned, which ensures maximum entropy and resistance to these attacks, as all possibilities are equally likely. Previous studies of graphical recognition based passwords have shown that user-chosen graphical

passwords are so vulnerable to dictionary attacks that user choice is expressly advised against (Davis, Monrose & Reiter, 2004). All password-based schemes are still subject to brute force attacks. Because of this, care must be taken to ensure that each authentication scheme in a study is equally resistant to them, designed with equal strength.

The strength of any password system lies with its associated password space. The larger the potential variety of password combinations (or password space) available for use, the more guesses will have to be attempted before there is success, and thus the more secure the system can be. In contrast to this, the principle of psychological acceptability states that authentication mechanisms must be designed for ease of use:

> "To the extent that the user's mental image of his protection goals matches the mechanisms he must use, mistakes will be minimized. If he must translate his image of his protection needs into a radically different specification language, he will make errors." (Saltzer & Schroeder, 1975)

In the context of passwords, the principle of Psychological Acceptability suggests that passwords should be easy enough for users to remember. Usability is thus critically important when designing a password system, in order that the system be accepted by its users and not circumvented (Adams & Sasse, 1999). In the context of passwords for authentication, usability is largely tied to memorability. These two factors, usability and security, are frequently at odds, and this dilemma is known as the usability / security tradeoff. Simple passwords are easy to remember and use, but offer little security. Complex passwords that include many letters, numbers and special characters are considered very secure, however they are much less memorable and thus more difficult to use.

The theoretical password space of a system is the set of all possible unique combinations allowed by that system. Theoretical password space can be calculated, and this is elaborated on

and demonstrated in the discussion that follows. However, a more meaningful evaluation of the password space associated with the security mechanism in question will reference its *effective* password space. The effective password space is the set of all possible password combinations that people may actually use, within the theoretical password space. For example, in text password systems people are almost certainly not going to choose "XzalCH49fQi5" due to its complexity. As previously mentioned in the graphical passwords section of this document, users of the PassFaces system tended to choose attractive faces of people sharing their race, and users of the Draw-a-Secret system tended to draw symmetric patterns to use during authentication. Weaknesses such as these allow attackers to narrow their password dictionaries in guessing-based attack methods, rendering their attacks increasingly successful. This study is principally concerned with ensuring consistent strength against brute force attacks, so we must ensure that the effective password space is the same as the theoretical password space, and the same in any conditions to be compared.

Table 1: Text password criteria and theoretical password space.

Description	Number of chars.	Length	Password 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

The size of any password space can be calculated by determining the number of total password variations that can exist within a set. For example, a bank account's PIN is created

from a set of 10 digits, of which the owner must select four, thus $10^4 = 10{,}000$ possible combinations. The strength of a given password space is typically expressed in bits, which can be calculated by taking the base two logarithm of the size of the password space. This method is common because computer memory stores information in binary digits (bits) consisting of only zeros and ones, and it would require that many bits to store a maximally complex password. In the case of the bank account PIN as seen in Table 1, we can calculate $\log_2(10^4) = 13$, expressed as 13 bits. This is because the number of possible passwords is the number of choices for each character (10), to the power of the length of the password (4), so the total number of passwords is 10,000. The base two logarithm of this number indicates the number of bits (13) necessary to store a number this large. The final example in Table 1 describes a password system allowing any character possible on a standard keyboard, including all lower case and upper case letters as well as numbers and special characters (95 in all). If a password in this system were eight characters long, the strength of this password would be measured as 53 bit strength. The other lines express additional possibilities.

Table 2: Potential Authentication Mechanisms and Theoretical Password Space.

Mechanism	Alphabet Name	# Of Items	Length / Clicks	# Of Combinations	Bit Strength
Word Recognition	All 6 Letter Words	15200	6	1.23328E+25	84
	Set of 96 Words	96	6	7.82758E+11	40
	Planned Recognition	26	6	308915776	28
Word Recall	All 6 Letter Words	15200	4	1.23328E+25	84
	Set of 96 Words	96	6	7.82758E+11	40
	Planned Word Recall	156	4	592240896	29
Letter Recall	Full Keyboard	95	13	5.13342E+25	86
	Full Keyboard	95	6	7.35092E+11	40
	Lower Case Only	26	6	308915776	28

To compare a textual recognition scheme with text recall schemes, the theoretical password spaces must be the same. Table 2 illustrates some of the possibilities. One of the scenarios (Word Recognition) might involve passwords that are sets of several words, each word being displayed on screen amongst a set of distracter words, available to be chosen by the user. Similar to the PassFaces scheme, users would be asked to recognize one word from a display of 26 total words, and to do this 6 times in a row. The space for this example is calculated in a similar manner as the traditional password schemes. Our condition will display 26 words on 6 sequential panels of words, and therefore will possess $\log_2 26^6 = 28$ bit security.

The second mechanism (Word Recall) would require users to remember a list of 4 whole words, which will serve as one password. These 4 words are taken from a set of 156 possible words, and so its password space is calculated as $\log_2 156^4$, thus exhibiting the same 28-bit security. The third example of text based authentication (Letter Recall) involves a password

composed of 6 random, lowercase letters which also possesses 28-bit strength, as seen in Table 2 above. As Table 2 shows, each of these schemes offer several potential theoretical password spaces, but it also shows that it is possible to configure different schemes with similar password spaces. Lastly, assigning the passwords used in any of these scenarios should prevent any disparity between the theoretical and effective password spaces.

Research Question

We wished to investigate the documented disparity between the effectiveness of recognition and recall, in the context of text-based password systems. By comparing participants' abilities to recall text-based passwords and their ability to recognize words as passwords, we were able to assess practical limitations of memory and the implications on user authentication. Furthermore, preserving a constant password space across all three conditions increased the validity of our observations.

When people have the opportunity to choose their passwords, they tend to create passwords that are as simple as possible, since those are most easily remembered. It is not clear if people are able to effectively remember passwords that are assigned to them, especially when the memory task is recognition. To summarize, the research question is: Can a recognition-based text password system facilitate authentication to a greater extent than traditional text passwords?

Experiment 1

Our study used a within-subjects design consisting of three experimental conditions. Each condition required participants to employ a different text-based authentication mechanism to log in and interact with web sites set up specifically for use in this experiment. The conditions used the three schemes described in Table 2, above. The first password type was a traditional six-

character random text password, composed of lowercase alphabetical characters (the letter recall condition). The second consisted of four assigned whole words, which when entered at the prompt served as one password (the word recall condition). The final condition was a cognometric graphical password system, except that rather than displaying a set of pictures for participants to select from in order to authenticate, they were shown panels of whole words which could be clicked in series to authenticate (the word recognition condition). Each of these conditions was implemented using 28-bit strength, as described in Table 2, above.

The word and letter recall conditions both represented authentication conditions involving pure recall, and the word recognition condition presented an opportunity to capitalize on recognition as retrieval mechanism. Both the word recall and word recognition conditions possessed the potential advantage of using whole words in passwords. This should allow users to process their passwords more deeply when attempting to memorize them because of the semantic meanings associated with words, which the letter recall condition does not allow (Craik & Lockhart, 1972).

The type of authentication used served as the independent variable (IV) in each of the planned analyses. To evaluate the hypotheses stated below, we needed to measure the length of time each type of password was remembered by each participant as the dependent variable (DV, Maximum Memory Time) by recording the amount of time between password creation (or reset) and the last successful login. If there were no resets, then the memory time would be the time between the beginning and end of participation, which was one week. We also measured the number of resets requested per participant per condition (Resets). Login efficiency was also measured as a DV, which was recorded as the time taken for each participant to authenticate successfully (Login Time). Lastly, the number of passwords that persist in memory for the duration of the study was recorded as a DV (Remembered Passwords).

Research Hypotheses

After having reviewed the theory behind the effectiveness of recognition and recall, and making some interpretation based on the data supporting graphical passwords, we were prepared to identify some hypotheses regarding the outcome of this experiment. Because recognition judgments have been described as more effective over lengthy periods of time, it is believed that the word recognition condition will result in significantly more memorable passwords than the two conditions relying on the process of pure recall, as stated in hypothesis one:

$H1_1$: There will be significantly greater memorability in the word recognition authentication system when compared to the recall-based (text entry required) authentication mechanisms, measured according to maximum memory time.

$H1_0$: There will be no significant difference in memorability when using the recognition-based system vs. the recall-based systems, based on the measure of maximum memory time.

For our second and third hypotheses, there was less certainty associated with each of the authentication methods. The three authentication methods are very different, creating a situation where novelty may play a role and the time required to type (or identify) the passwords will cause an unknown influence. Because of this, the second and third hypotheses are non-directional.

$H2_1$: There will be a significant difference in the number of password resets initiated for each type of authentication.

$H2_0$: There will not be a significant difference in the number of password resets initiated for each type of authentication.

$H3_1$: There will be a significant difference in time required to log in across authentication types.

$H3_0$: There will not be significant difference in time required to log in across authentication types.

In an attempt to measure the simple effectiveness of each authentication mechanism, we compared the number of passwords that are remembered correctly at the end of their participation. As with hypothesis one, it was expected that recognition would be superior to recall, as outlined below:

$H4_1$: There will be a significantly greater number of passwords remembered for the duration of the study in the Word Recognition condition than in either of the two recall related conditions.

$H4_0$: There will not be a significant difference in password memorability across each of the authentication mechanisms.

Method

Participants.

Participants recruited for this experiment were individuals who made regular use of the Internet and web sites that require authentication. Participation was restricted to those who do not have any serious visual or memory related impairments that may have affected the outcome of this investigation. The participants for this study consisted mainly of university students, and young adults who were compensated for their time either in the form of twenty dollars, or two bonus percentage points in the undergraduate Psychology course in which they were enrolled. They were recruited for the study via Carleton's SONA system, posters spread throughout the

campus (visible in Appendix G), as well as word-of-mouth. Potential participants who did not have access to SONA were able to register via e-mail.

Materials.

To administer this experiment we created three websites that required authentication in order to view and contribute content (screenshots available in Appendix H). Automated reminders were sent to our participants at regular intervals (example in Appendix G), asking them to log in by entering their three assigned passwords at the prompts as visible in Figures 4 and 5 below.

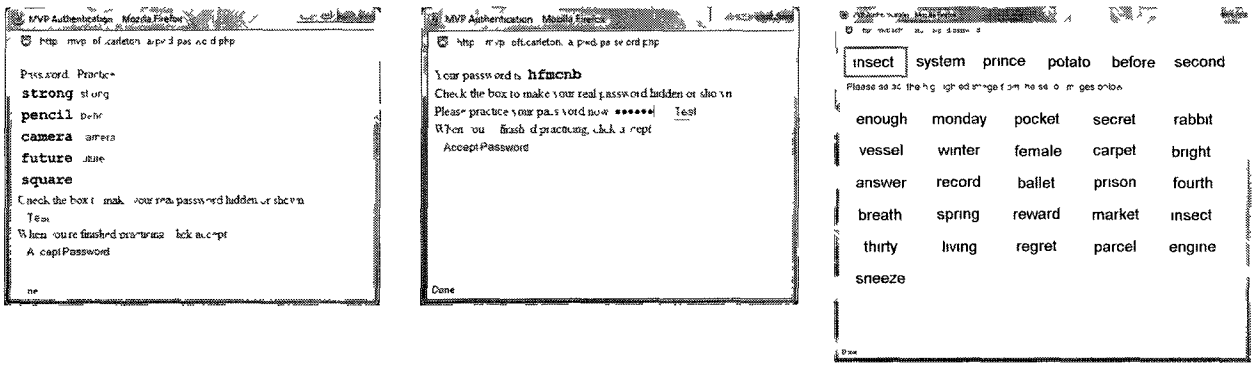


Figure 4: Examples of the registration screens for the Word Recall, Letter Recall and Recognition conditions, respectively.

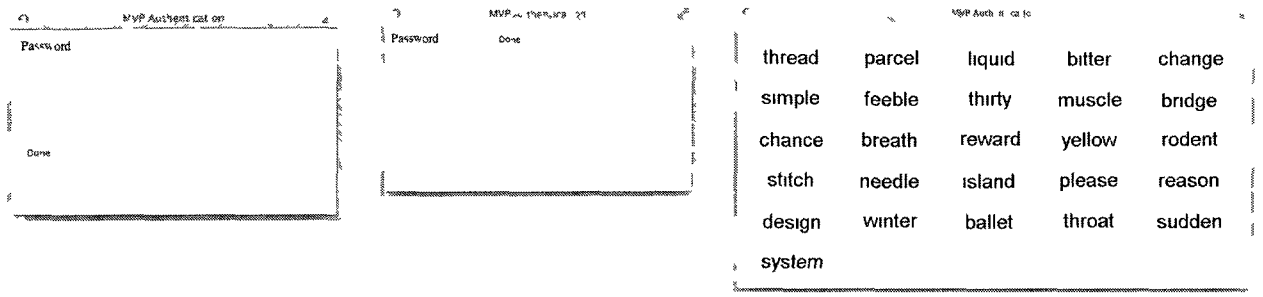


Figure 5: Examples of the login screens for the Word Recall, Letter Recall and Recognition conditions, respectively.

The set of words used to generate the passwords we assigned in the word recognition and word recall conditions were selected from Ogden's Basic and International word lists (Bauer, 2008). Our words were chosen from Ogden's lists in order to create a selection of words that is representative of daily language (the selection is shown in Appendix I). By creating passwords from words widely recognized as central to an understanding of the English language, we hoped to ensure that all participants were familiar with them, further controlling for individual differences among participants. The letters used in the letter recall condition consisted of the 26 letters (lower case) found in the English language.

We administered informed consent and debriefing forms, as required by Carleton University's Ethics Committee for Psychological Research, which can be seen in Appendices A and E. We also used forms to collect demographic information as well as pre-test and post-test opinions and feedback (visible in Appendices B, C & D). Participants needed only access to a computer and a valid e-mail address to participate.

Apparatus.

Participants used personal computers (PCs) with Internet access in our lab and at home to authenticate at each of the three sites created for use in this study. The three websites had each been outfitted with a password protection scheme according to our conditions of interest. In our lab they used PCs operating with Microsoft Windows 7 and browsed the Internet using Internet Explorer. Between lab sessions, the participants were free to use whatever combination of computer and Internet browser they preferred. The MVP authentication framework was used to facilitate account management and automated participant reminders (Chiasson, Deschamps, Stobert, Hlywa, Freitas, Machado, Chan, & Biddle, 2009).

Procedure.

Phase 1.

Participants arrived at the lab at a time previously agreed upon. At arrival, participants were given an explanation of the experiment, and told that they would be able to withdraw from the experiment at any time without penalty. A consent form was then provided for them to read and sign, if they agreed, before the experiment commenced.

After providing their consent, the participants were shown to a computer and given a simple introductory questionnaire, which gathered demographic information. The participants were then given their first password and asked to sign-in to a web site, and then sign out. They were then shown to the other two sites, given those passwords and asked to learn them and authenticate. They were then asked to authenticate to each of those same sites again, to demonstrate that the password had been memorized. If they were unable to reproduce their password and log in successfully, they were shown their password again and encouraged to login

until they could do so without requiring any help, to ensure that they had memorized their passwords before the lab session was concluded.

Note that each password was of a different type, and each site was distinct. Counter balancing was employed in this study to control for order effects so that the participants did not all see the same schemes in the same order. This served to protect against the possibility of order effects influencing the data across each authentication condition.

Once all three passwords had been memorized, an appointment was scheduled for the second lab session approximately one week later. They were then encouraged to login from home and told to expect two notification e-mails, and the session was then concluded.

Phase 2.

Over the period of one week, while at home, participants received two reminder e-mails asking that they log into each of the three sites for which they were assigned passwords and to contribute to the content of the site. The participants were free to do this on any computer that they could access. All actions taken with regards to authentication were logged on the web servers, so the time required to log in, number of successful as well as failed attempts, and the time and number of password resets were recorded for later analysis.

Phase 3.

At the second scheduled appointment, which took place at an agreed upon time approximately one week after the first, the participants arrived at the lab and were greeted by a researcher. They were then shown to a computer, asked to log into each of the sites and add a written entry to each of the web sites one last time. When they had finished, they were given a questionnaire related to the password schemes and their experiences throughout the study. Upon

submitting the questionnaire, they were handed the debriefing form and encouraged to ask any questions or voice any concerns they may have had. They were then compensated and thanked for their time, and their participation in the experiment was then concluded.

Analysis Plan

A system was put in place that recorded all authentication-related activity that occurred on the three web sites created for the purposes of this study. We captured this data during participants' first and second visits to the lab and throughout their week long trial. We sent out two reminder e-mails during this between-visits period at intervals of two and four days following their initial lab session, inviting them to attempt to authenticate at least twice from a natural setting. All login names and passwords were assigned to participants, thus absolutely no personal data were used on the web sites. From the logs of usage for these sites we were able to extract data about each users' time spent logging in, number of trials per attempt, the number of times a password was recovered, and the number of successful or failed attempts. This information allowed us to address each of the hypotheses below. In all cases, we used an alpha level of 0.05, and limits of -2.0 and 2.0 for skewness and kurtosis when performing tests dependent on normality.

Hypothesis One.

$H1_1$: There will be a significantly greater memorability in the word recognition authentication system when compared to the recall-based (text entry required) authentication mechanisms, measured according to maximum memory time.

$H1_0$: There will be no significant difference in memorability when using the recognition-based system vs. the recall-based systems, based on the measure of maximum memory time.

To test this hypothesis, two paired T-tests were to be conducted, to compare the recognition condition with each of the recall conditions. Authentication mechanism was to serve as the independent variable, and maximum memory time was to be the dependent variable. In addition, a repeated-measures ANOVA was to be performed to identify any overall significant difference.

Hypothesis Two.

$H2_1$: There will be a significant difference in the number of password resets initiated for each type of authentication.

$H2_0$: There will not be a significant difference in the number of password resets initiated for each type of authentication.

To test this hypothesis a repeated-measures ANOVA was to be conducted on the mean number of resets used in each condition. Authentication mechanism was to serve as the independent variable, and the mean number of password resets as the dependent variable.

Hypothesis Three.

$H3_1$: There will be a significant difference in time required to log in across authentication types.

$H3_0$: There will not be significant difference in time required to log in across authentication types.

To test this hypothesis we were to conduct a repeated-measures ANOVA between the distributions of time required to authenticate for each form of authentication. The authentication

mechanism was to be used as the independent variable and the mean login time was to be used as the dependent variable in this analysis.

Hypotheses two and three require the use of repeated-measures ANOVA tests. These are based on assumptions regarding the normality of the data, which was to be verified. While ANOVA tests are robust to violations of these assumptions, especially with sufficient sample sizes, non-parametric analyses or transformations were to be performed if necessary.

Hypothesis Four.

$H4_1$: There will be a significant difference in the number of passwords remembered for the duration of the study, across each of the experimental conditions.

$H4_0$: There will not be a significant difference in the number of passwords remembered for the duration of the study, across each of the experimental conditions.

To test for a significant difference in number of persistent passwords amongst the three conditions, a chi-squared analysis was to be performed. The authentication mechanism used was to serve as the independent variable and the number of passwords used successfully throughout the experiment will serve as the dependent variable in this case.

Analyses of Interest.

With the information we will have collected in an effort to respond to each of the aforementioned hypotheses, we will also be able to review whether or not there are other factors influencing the memorability / usability of each of the forms of passwords in use.

Since the password stimuli were composed of whole English words in two of the three conditions, it is possible that a person's first language may influence their performance in this experiment, specifically in the recognition and word recall conditions. By requesting that

participants disclose their first language on the demographic information questionnaire, we were able to contrast that information with their performance data to evaluate for any potential effects.

A gender bias may be present within the results. Typically there are not, as men and women are thought equally capable of recognizing and recalling text, however, one security related study (Chiasson, Forget, Stobert, Biddle & Van Oorschot, 2009) did report a difference in the ability for men and women to authenticate when using graphical passwords. We did not precisely gender-balance the experiment, but did plan to explore the effects of gender.

Each participant was to be assigned three novel passwords, and was expected to use them several times over a period of a week or more. Barring attrition, practice effects may have come into play. Participants may either have experienced greater difficulty recalling or recognizing their passwords over time, or their performance may have improved, as their passwords will grow increasingly familiar. To assess this we planned to evaluate performance over time on successful login attempts, for each participant and in each condition.

Results

Participants.

We recruited 38 participants for our principal investigation, and everyone completed session 1. Upon analysis, the data for participants 8 and 17 were omitted as they never successfully logged in, or reset their passwords following the initial laboratory session; Participants 37 and 38 were added in their place. Participant 1 was unable to complete the initial survey due to a power outage on the date of their appointment, but all other data originating from this participant was valid, and included in our results, so our dataset consists of a full week of observations on 36 participants.

The 36 participants comprised 15 males and 21 females, with a mean of 29.8 years of age. Twenty-five of them had a social science related background and seven reported a natural science or engineering related background, with the remaining four participants choosing not to disclose. Participants showed an average of 3.87 years of post-secondary education. Twenty-nine (or 80.6%) of respondents had English as a first language, and the other seven (or 19.4%) spoke English as a second language. When asked to rate their computer skills on a scale from 1 to 10, where one meant "novice" and ten meant "expert", this sample's mean was 7.08, the median response was 8, and one person rated themself a 3, which was the lowest response. The vast majority of participants (91.7%) use the Internet daily, and the others all reported using it several times per week. Lastly, while everyone participated in the two lab sessions, participation dropped by about one third for the first attempt from home, and roughly half of the participants made an attempt after receiving the second e-mail notification.

Descriptive Statistics.

Table 3: Descriptive Statistics For All Authentication Conditions

	N	Min.	Max.	Mean	Median	Std. Dev.	Skewness		Kurtosis	
							Statistic	Std. Error	Statistic	Std. Error
LRSuccess	36	0	6	2.86	3	1.2	0.177	0.393	0.907	0.768
RecSuccess	36	0	5	2.36	2	1.18	0.354	0.393	-0.085	0.768
WRSuccess	36	0	6	2.61	3	1.4	-0.175	0.393	0.21	0.768
LRFail	36	0	14	1.31	0	2.8	3.328	0.393	12.495	0.768
RecFail	36	0	9	1.31	0	2.2	2.02	0.393	3.706	0.768
WRFail	36	0	17	2.86	1	4.15	2.029	0.393	4.299	0.768
LRReset	36	0	1	0.08	0	0.28	3.148	0.393	8.371	0.768
RecReset	36	0	1	0.03	0	0.17	6	0.393	36	0.768
WRReset	36	0	2	0.28	0	0.51	1.687	0.393	2.164	0.768
LRLogTime	36	4.5	25.2	9.36	8.16	4.77	2.07	0.393	4.449	0.768
RecLogTime	36	22	181	58.68	50.67	32.79	2.05	0.393	5.353	0.768
WRLogTime	35	6	109.25	22.93	15.83	20.56	3.02	0.398	9.831	0.778
LRMaxMem	36	0	202.45	153.17	167.04	49.15	-1.927	0.393	4.527	0.768
RecMaxmem	36	0.19	201.53	161.05	166.84	37.6	-2.37	0.393	10.199	0.768
WRMaxmem	36	0.06	199.07	120.5	164.56	74.89	-0.83	0.393	-1.099	0.768

Table 3 Shows the descriptive statistics associated with each of the three authentication conditions present in this study. In the interest of brevity, the letter recall condition has been abbreviated as "LR", recognition as "Rec" and word recall as "WR". This table allows for an overall view of the statistics related to login successes, failures and resets, along with the time taken to login and maximum memory times for each condition.

Hypothesis One.

$H1_1$: There will be a significantly greater memorability in the word recognition authentication system when compared to the recall-based (text entry required) authentication mechanisms, measured according to maximum memory time.

$H1_0$: There will be no significant difference in memorability when using the recognition-based system vs. the recall-based systems, based on the measure of maximum memory time.

To test this hypothesis, a repeated-measures ANOVA was to be performed to identify any overall significant difference and paired T-tests were to be conducted to isolate any significant differences between conditions. Authentication mechanism served as the independent variable, and maximum memory time was the dependent variable. The maximum possible value for memory time is about 200 hours, because participants were enlisted for a period ranging from six to eight days. Before we could perform an ANOVA it was necessary to verify the normality of the memory time distributions, which are graphed in Figure 6.

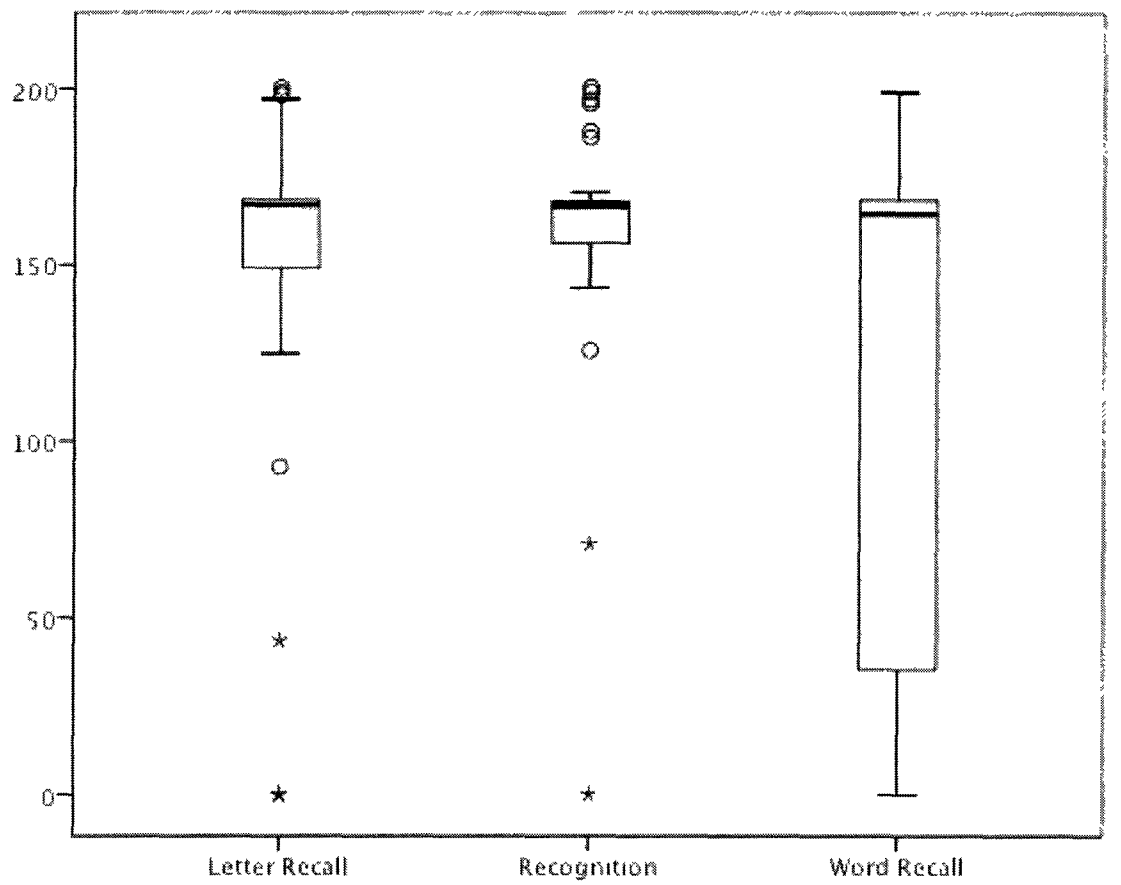


Figure 6: Boxplots of memory persistence in each authentication condition

Because the observed results for maximum memory time do not meet the assumptions of normality (skewness measurements of -1.93, -2.37, and -0.83, all with S.E of 0.393; kurtosis measurements of 4.53, 10.20 and -1.09, with S.E. of 0.768, across the letter recall, recognition and word recall conditions respectively), the planned ANOVA could not be performed. For these distributions to have been normal, most people would have had to have forgotten their passwords by the end of the study. Previous research in this area has yielded normal results, and they were anticipated in this study. While normality would allow for stronger statistical results, the fact that many participants remembered their passwords for the duration of their participation

speaks to the usability of the conditions in question. The non-parametric test known as Friedman's test was used in its place, and the distributions used in this test are described in Table 4. Friedman's test is a non-parametric (and inherently weaker) version of an ANOVA test like Kruskal-Wallis, however, Friedman's test accounts for the within-subjects design used in this experiment, which accommodates individual differences.

Table 4: Maximum Memory Time Descriptive Statistics

							Skewness		Kurtosis	
	N	Min.	Max.	Mean	Median	Std. Dev.	Statistic	Std. Error	Statistic	Std. Error
Letter Recall	36	0	202.45	153.17	167.04	49.15	-1.927	0.393	4.527	0.768
Recog.	36	0.19	201.53	161.05	166.84	37.6	-2.37	0.393	10.199	0.768
Word Recall	36	0.06	199.07	120.5	164.56	74.89	-0.83	0.393	-1.099	0.768

Friedman's test showed no significant differences between the conditions in this study ($\chi 2 = 1.167$, $p = 0.558$). Therefore recognition did not prove significantly more memorable in this case. Because there was no significant difference identified, no further tests were conducted. Therefore no support for hypothesis one was found. As a non-parametric test based on ordinality Friedman's test does not address skewness. The obvious skewness of the distribution of the word recall condition is quite striking in this case, and indicates that a larger number of people had difficulty remembering their passwords for a comparable period of time.

Hypothesis Two.

$H2_1$: There will be a significant difference in the number of password resets initiated for each type of authentication.

$H2_0$: There will not be a significant difference in the number of password resets initiated for each type of authentication.

To test this hypothesis a repeated-measures ANOVA was to be conducted on the number of resets used in each condition. Authentication mechanism served as the independent variable, and the mean number of password resets served as the dependent variable.

Because the observed results for password resets did not meet the assumptions of normality (skewness measurements of 3.15, 6, and 1.69, all with S.E of 0.393; kurtosis measurements of 8.37, 36 and 2.16, with S.E. of 0.768, across the letter recall, recognition and word recall conditions respectively), the planned ANOVA could not be performed. Participants were allowed to reset their passwords as many times as they chose throughout their participation, however, nobody reset their passwords more than twice per authentication condition, as outlined in Table 5. Figure 7 displays the distributions of reset attempts by condition, and the skewness and kurtosis measurements for each distribution of resets were all greater than two. The non-parametric test known as Friedman's test has been used in its place.

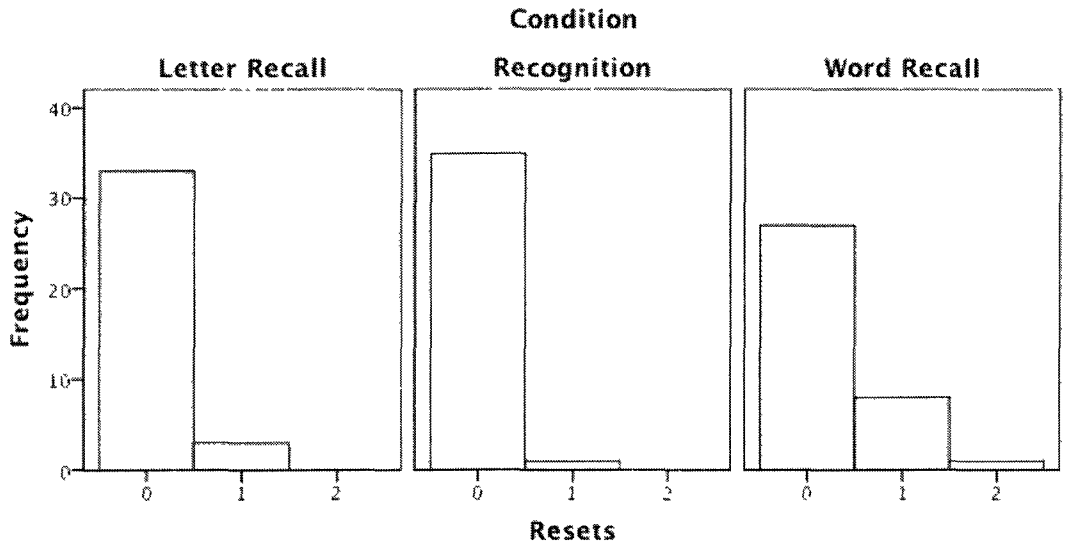


Figure 7: Histograms of reset frequency per authentication condition

Table 5: Password Resets Descriptive Statistics

	N	Min.	Max.	Mean	Median	Std. Dev.	Skewness		Kurtosis	
							Statistic	Std. Error	Statistic	Std. Error
Letter Recall	36	0	1	0.08	0	0.28	3.148	0.393	8.371	0.768
Recog.	36	0	1	0.03	0	0.17	6	0.393	36	0.768
Word Recall	36	0	2	0.28	0	0.51	1.687	0.393	2.164	0.768

Friedman's test verified the presence of a significant difference in number of password resets requested between the authentication conditions ($\chi 2 = 9.455$, $p = 0.009$). Due to violation of the assumptions required for normality, the investigation used Wilcoxon paired tests for post hoc analysis.

After applying Bonferroni corrections, the recognition condition had significantly fewer resets than the word recall condition ($Z = 2.496$, $p = 0.039$), however the difference between the letter recall and word recall conditions was not significant ($Z = 2.111$, $p = 0.105$), and the

difference between the recognition and letter recall conditions also showed no significance ($Z = -1.000$, $p = 0.951$). Hypothesis two has been supported by these findings.

Hypothesis Three.

$H3_1$: There will be a significant difference in time required to log in across authentication types.

$H3_0$: There will not be significant difference in time required to log in across authentication types.

To test this hypothesis we were to conduct a repeated-measures ANOVA between the distributions of time required to authenticate for each form of authentication. Authentication mechanism was used as the independent variable and the average login time was the dependent variable in this analysis.

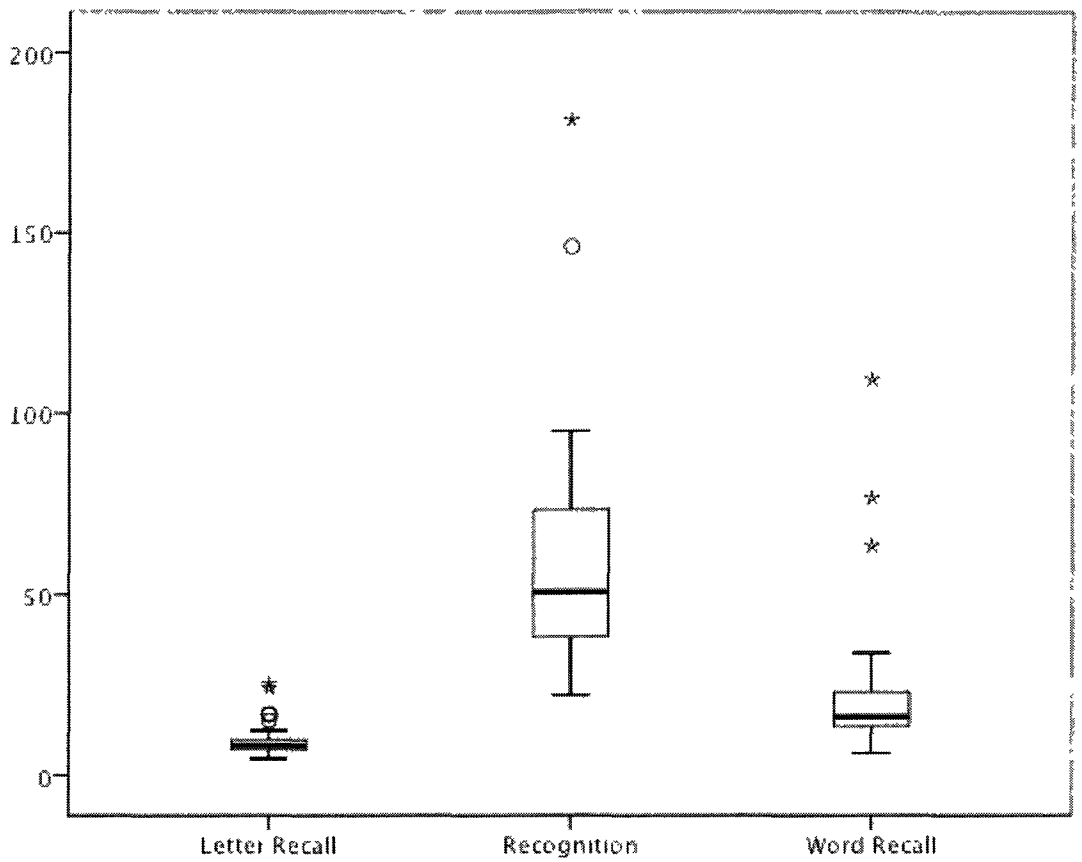


Figure 8: Boxplots of performance: login time per authentication condition

When reviewing the observations for login time associated with each condition (as shown in Figure 8), it was revealed that the distributions were not normal. Similar to the findings for password resets, the skewness and kurtosis measurements for login times associated with each authentication condition were all greater than 2.0.

Because the observed results did not meet the assumptions of normality, the planned ANOVA could not be performed. The descriptive statistics associated with this investigation, including skewness and kurtosis are shown in Table 6. The non-parametric test known as Friedman's test was used in its place.

Table 6: Login Time Descriptive Statistics

	N	Min.	Max.	Mean	Median	Std. Dev.	Skewness		Kurtosis	
							Statistic	Std. Error	Statistic	Std. Error
Letter Recall	36	4.5	25.2	9.36	8.16	4.77	2.07	0.393	4.449	0.768
Recog.	36	22	181	58.68	50.67	32.79	2.05	0.393	5.353	0.768
Word Recall	36	6	109.25	22.93	15.83	20.56	3.02	0.398	9.831	0.778

Friedman's test was significant in this case and deserving of further investigation ($\chi 2 = 60.743$, $p < 0.001$). Because the data did meet the assumptions of normality, we continued to investigate using three Wilcoxon paired tests.

All differences were significant in this evaluation. After applying Bonferroni corrections, the recognition condition was significantly different from the letter recall condition ($Z = -5.160$, $p < 0.001$) and the word recall condition ($Z = -4.769$, $p < 0.001$), and there was also a significant difference between the letter and word recall conditions ($Z = -4.845$, $p < 0.001$). Therefore the letter recall authentication mechanism handily outperformed the other two, in terms of time required to login successfully. These differences support hypothesis three.

Hypothesis Four.

$H4_1$: There will be a significant difference in the number of passwords remembered for the duration of the study, across each of the experimental conditions.

$H4_0$: There will not be a significant difference in the number of passwords remembered for the duration of the study, across each of the experimental conditions.

To test for a significant difference in number of persistent passwords amongst the three conditions, a Chi-squared analysis was performed. The authentication mechanism used served as

the independent variable and the number of passwords used successfully throughout the experiment served as the dependent variable in this case. Figure 10 displays the number of participants who remembered their passwords for each condition throughout the duration of the study, and as well as counts of the passwords that were not remembered throughout.

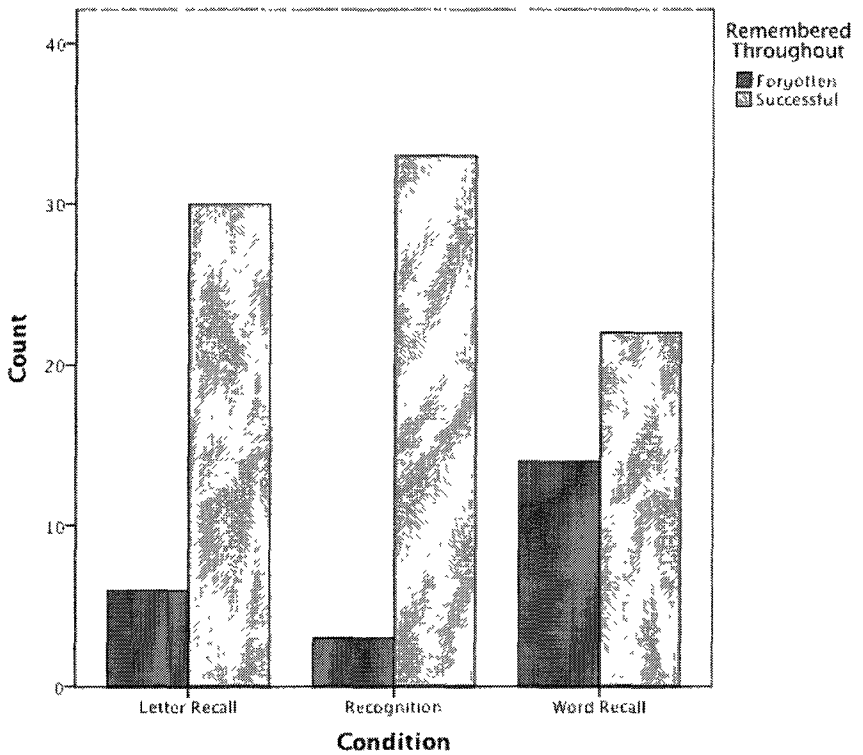


Figure 10: Bar graph of passwords remembered throughout the study per condition

Table 7: Crosstabulation of MemToEnd by Condition

			Letter Recall	Recognition	Word Recall	Total
Remembered Throughout	No	Count	6	3	14	23
		Expected Count	7.7	7.7	7.7	23
	Yes	Count	30	33	22	85
		Expected Count	28.3	28.3	28.3	85
Total		Count	36	36	36	108
		Expected Count	36	36	36	108

Table 7 outlines the counts for the chi-squared test, which indicated a significant difference in participants' abilities to remember their passwords for the duration of the experiment ($\chi 2 = 10.717$, $p = 0.005$). We then continued to explore the differences using three chi-squared tests pairing each authentication condition with the others.

Having performed the three chi-squared post-hoc tests and applying Bonferroni corrections, we found no significant difference between the letter and word recall conditions ($\chi 2 = 4.431$, $p = 0.105$), nor the recognition and letter recall conditions ($\chi 2 = 1.143$, $p = 0.885$), though the recognition and word recall conditions showed a significant difference ($\chi 2 = 9.318$, $p = 0.006$). Hypothesis four is supported by these results.

Summary of Hypothesis Tests.

Reflecting upon the results of these hypothesis tests, there is a clearer picture of the influence of recognition on password usability. The outcome of hypothesis one revealed that there were no significant differences between authentication conditions in terms of maximum memory time.

When considering the number of password resets that occurred in each condition (per hypothesis 2), the recognition condition performed significantly better than the word recall condition, though not significantly better than the letter recall condition. The number of passwords that were remembered for the duration of the study is another important usability metric for which we could track the results. The fourth hypothesis revealed findings very similar to those of the second hypothesis, where the recognition condition enhances the ability of people to remember the passwords they were assigned. The recognition condition produced significantly

more passwords remembered for the study's duration than the word recall condition, though not significantly different from the letter recall condition.

The third hypothesis test investigated a usability metric regarding performance time. Among the three authentication conditions, the recognition condition performed most poorly, requiring by far the greatest amount of time for participants to login successfully. Overall differences in this test were significant, with the letter recall condition permitting faster login times than the other two conditions.

The three memorability related metrics tested revealed either no significant difference, or simply a difference between recognition and word recall, but not letter recall. The time required to log in was significantly different for all pairings, and notably worst in the recognition condition. Further differences were investigated and reported in the analyses of interest or questionnaire results sections, below.

Analyses of interest.

Performance over time.

Upon completion of the study it was also possible to evaluate the results over time. Participants were sent e-mails requesting that they visit each of the three sites at intervals of two and four days after their first lab session, and on the sixth day they were sent another e-mail reminding them of their appointments for participation in the second lab session. Using these notifications as logical breaks in time, the website activity was grouped into four periods. The data for time period zero accounts for days zero and one, time two represents days two and three, time three represents days four and five, and the last period represents day six and beyond.

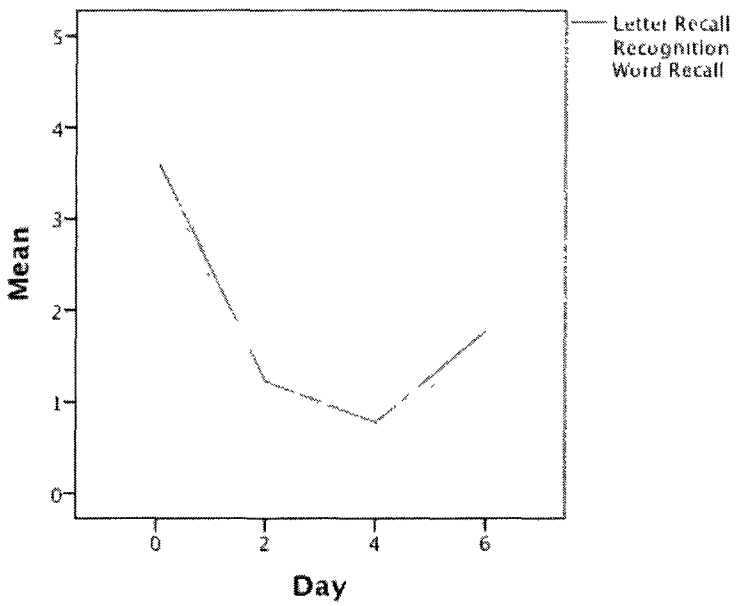


Figure 11: Login attempts per authentication period

Figure 11 depicts the mean number of trials made in each time period, for each authentication condition. Participants performed the greatest number of attempts when learning their newly assigned passwords, followed by reduced participation when away from the lab

throughout time periods two and four, and more attempts in the final time period. All conditions showed the same pattern regarding number of attempts over time.

It bears reporting that participation outside of the lab was noticeably reduced. Upon receiving the first task notification e-mail, only 24 participants in the letter recall condition, 23 in the recognition condition, and 22 in the word recall condition took any action on the websites used in this study, a reduction in participation of approximately one third. After the second e-mail notification, the participation rate was worse still, with only 17, 15 and 18 people taking action in the letter recall, recognition and word recall conditions respectively. Clearly our attempt at increasing the ecological validity of this experiment has an associated cost.

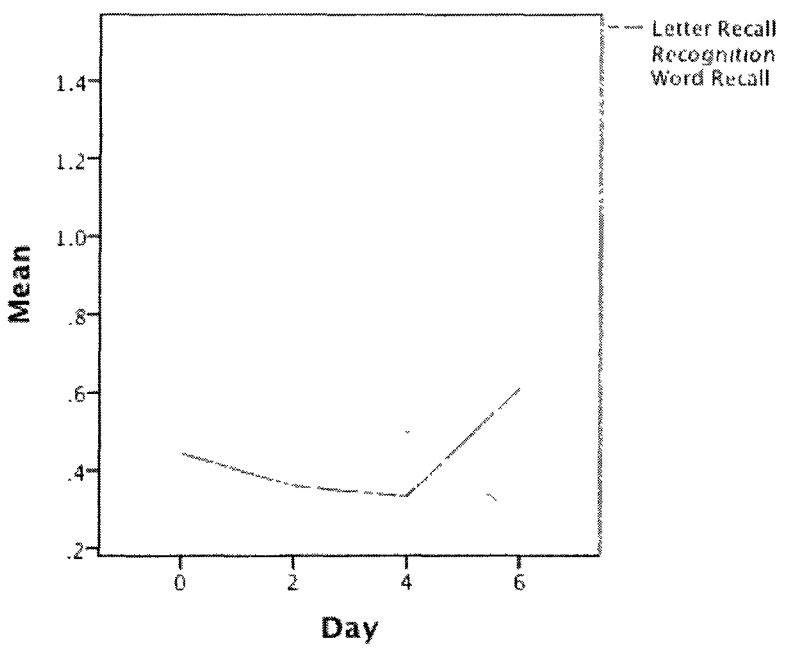


Figure 12: Mean failed authentication attempts per authentication period

The difference between conditions becomes more apparent in Figure 12, which plots failed attempts for each condition over the four time periods. The pattern for the word recall condition seemed notably different than that of the other two. The elevated number of failed

attempts in the initial and final time periods may be due to the learning phase and the passage of time. The reduction in failed attempts between the beginning and end of the study could be caused either by participants becoming more familiar with their passwords, or because they neglected to participate at all in a given period of time.

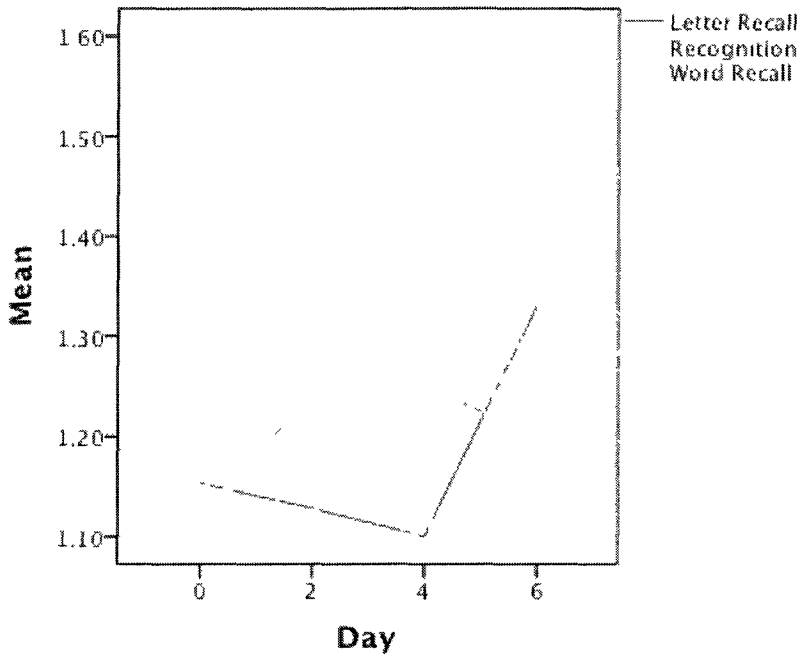


Figure 13: Mean number of attempts per successful login, by authentication period.

Upon viewing the number of login attempts per successful authentication as displayed in Figure 13, the distinction between the recognition condition and the two recall conditions becomes apparent once again. Wilcoxon paired tests were performed on each condition, testing the difference in number of attempts per successful login between the initial and final time periods. No single condition witnessed a significant change over time and Friedman's tests revealed no significant difference between the conditions at any period within the experiment.

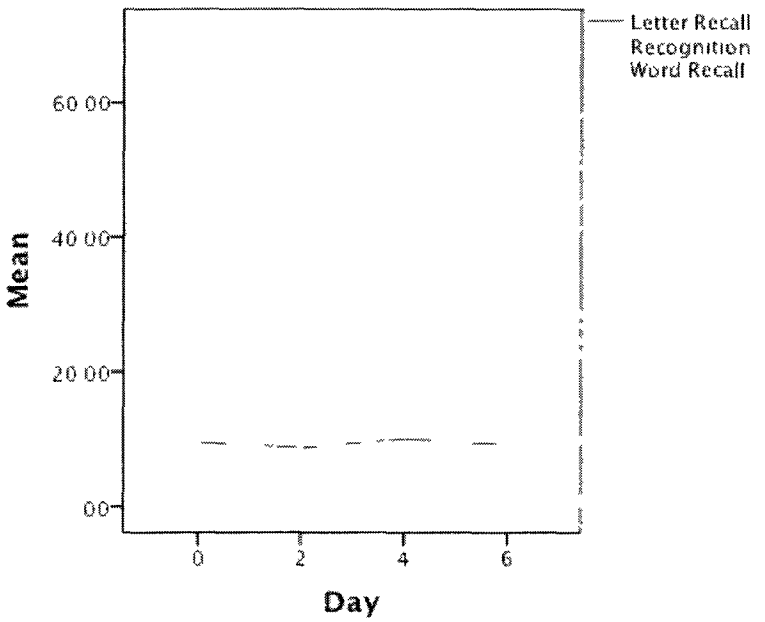


Figure 14: Mean login times per authentication condition over time

Figure 14 shows mean login times for successful attempts in each of the authentication conditions over time. There were no significant differences present in any one condition over time. The constant rise in the recognition condition initially caused some concern, however, it was not significant. It is conceivable that participants may have become more relaxed with their passwords over time. Alternatively, as the number of failed attempts stayed constant over time, perhaps the participants were trying harder and harder as time passed. All differences between conditions were significant at every time period in this investigation, meaning that throughout the experiment, the letter recall condition required the least time to login, followed by the word recall condition and then the recognition condition which on average required at least double the time to authenticate than using the letter recall condition.

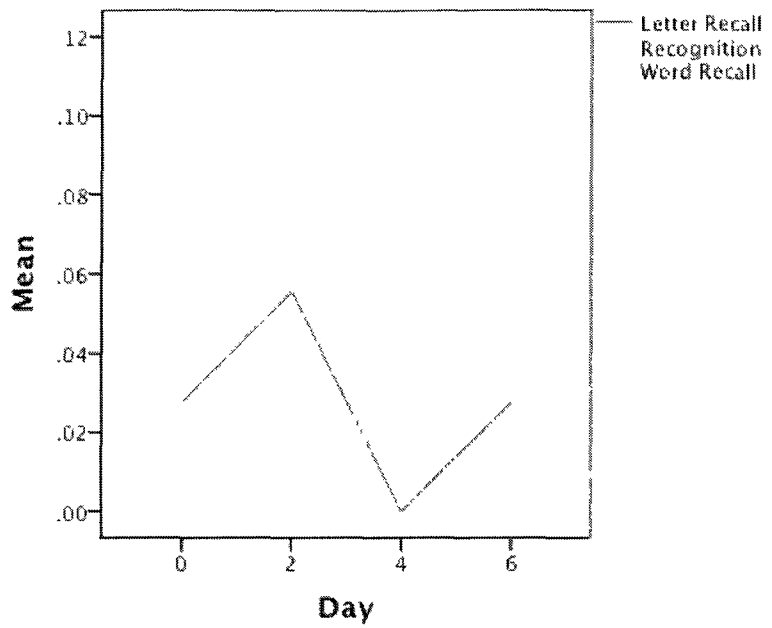


Figure 15: Mean password resets per condition by time period.

Kruskal-Wallis tests revealed that there was no significant change in any of the conditions between the first and last sessions. Friedman's test revealed no significant difference at the initial time period ($\chi 2 = 5.200$, $p = 0.074$), no significant differences between the conditions during the two intermediate periods between lab sessions, but a significant difference was present in the final session ($\chi 2 = 6.500$, $p = 0.039$). Investigating this difference using Bonferroni corrected post-hoc Wilcoxon tests revealed no significant differences between pairs of the conditions.

Influence of demographic factors.

During the initial lab session, immediately following the informed consent form, we administered a participant information survey that captured various demographic factors. Pursuant to the analyses of interest mentioned earlier, we looked for distinctions between participant categories, including age, gender, first language (English vs. other), field of study (social science related, or natural science and engineering), and years of post secondary education, with outcome variables, namely: resets, time to login, maximum memory time, and passwords remembered for the duration of the study.

These calculations revealed a significant correlation between age and login time ($r = 0.362$, $p < 0.001$). Further investigation then revealed that age was significantly correlated with login time in the context of the letter recall condition ($r = 0.494$, $p = 0.002$), and not significantly with the word recall condition ($r = 0.253$, $p = 0.143$) or the recognition condition ($r = 0.289$, $p = 0.002$). This bodes well for the recognition condition, in that aging will not affect our ability to authenticate as we age. It is however important to recall that the recognition condition required the greatest amount of time to authenticate successfully, by far.

That said, there were no other significant distinctions between the remaining demographic variables (gender, first language, field of study or years of post-secondary education) and any of the outcome variables (password resets, login time, maximum memory time and passwords remembered for the duration of the study).

Questionnaire Results

As part of this study's routine, participants were asked to fill out a demographics related questionnaire, and two surveys. The first was submitted at the end of the initial laboratory visit, and the second was administered at the end of the second laboratory visit, upon completion of their participation. The first survey asked about their password habits, and gathered some initial impressions of the password mechanisms used in this study. The second followed up on their opinions, and asked for comments related to their participation in the experiment. We can now explore some of those findings. General perceptions are presented first, followed by inquiries comparing the different schemes.

All of the passwords generated for use in this study were assigned to the participants, meaning they had no choice and no input into the content of their passwords. The study was designed in this manner to ensure maximum entropy among the passwords and comparable password spaces across conditions. However this is unlike our usual passwords, and we thought to ask participants about it by having them rate the statement "being assigned my passwords is frustrating" on a scale from one to ten, where one meant "strongly disagree" and ten meant "strongly agree", and the observations are shown in Figure 17.

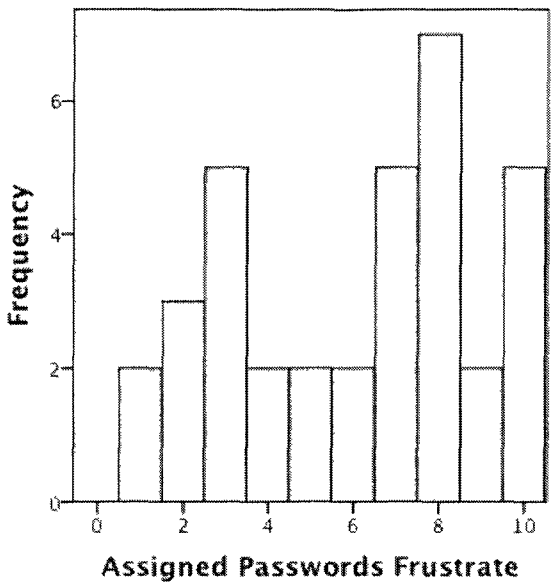


Figure 17: Histogram of ratings of frustration due to password assignment

The mean response was 6.06, with a standard deviation of 2.89. The graph shows a certain level of bimodality. This is to say that many participants either liked or hated the assignment of their passwords. The high level of frustration indicates that great care must be exercised in the design of future authentication mechanisms where password assignment may be considered.

Next, participants were asked to rate the likelihood that they would be able to remember their three passwords for the duration of the study on a scale from 1 to 10, where one meant "strongly disagree" and ten meant "strongly agree". The results are shown in Figure 18.

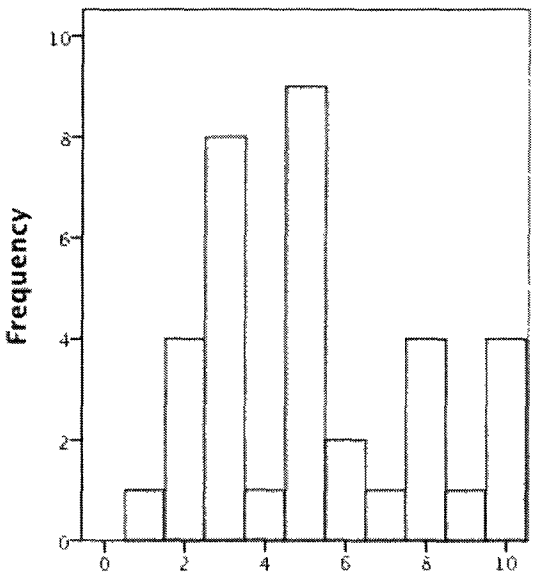


Figure 18: Histogram of participant-projected password memorability

The distribution of responses had a mean of 5.2 and a standard deviation of 2.67. This does not reveal a consensus of confidence or doubt, as much as a sense of uncertainty. Participants were also asked to estimate the number of websites they use that require authentication in order to use some of the functionality. The distribution of responses showed a minimum of 3 sites, a maximum of 50, and the mean response was 12.94 with a standard deviation of 9.57.

Password reuse was an obvious coping strategy among participants, with 88.9% of them reporting using the same passwords in different authentication systems. When asked about their criteria for selecting passwords in new systems, 71.4% said they chose passwords that were easy to remember, 60% reported using passwords that were the same as another, less than half (42.9%) included "difficult to guess" in their criteria, and 19.4% listed another criteria such as "similar to other passwords, but not the same" or the names of memorable people, places, dates or events. No one used passwords that were specifically suggested by the system, like "Ex@mple5".

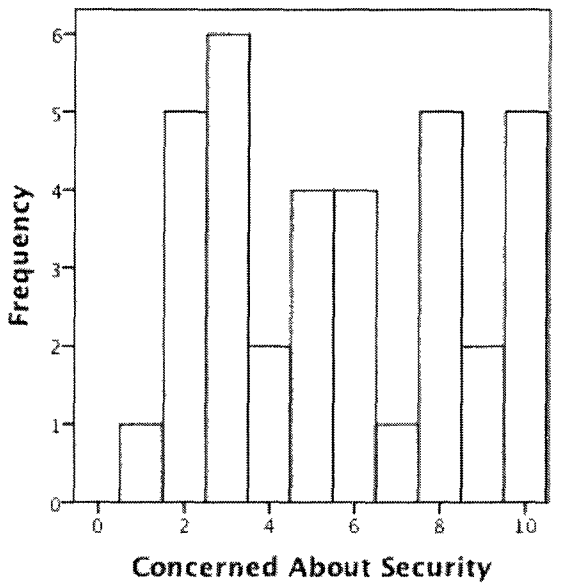


Figure 19: Histogram showing participants' concern for password security.

The histogram included as Figure 19 shows fairly evenly distributed levels of concern, with a mean response of 5.6 and an associated standard deviation of 2.89. Nonetheless it is somewhat encouraging that more people were "very concerned" than the one person who was "not at all concerned".

Participants were also asked a variety of questions, which allowed comparison of the three conditions. As part of the initial survey, one question asked respondents to rate the statement "passwords of type X are secure" on a scale from one to ten, where one meant "strongly disagree", ten meant "strongly agree", and each of the three different authentication types was substituted for "type X".

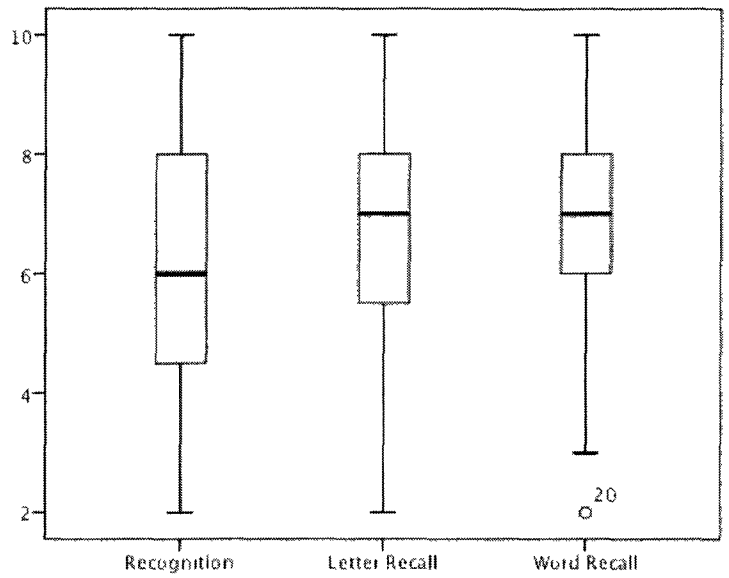


Figure 20: Boxplots of the security ratings for each authentication condition.

The next graph shows the distribution of responses when participants were asked to rate their concern for password security on a scale from 1 to 10, where one meant "not at all concerned" and ten meant "very concerned". Upon reviewing Figure 20 it became apparent that participants' perceptions of the security of the different authentication conditions were relatively similar. The difference was determined to be insignificant using a Kruskal-Wallis test ($\chi 2 = 1.579$, $p = 0.454$).

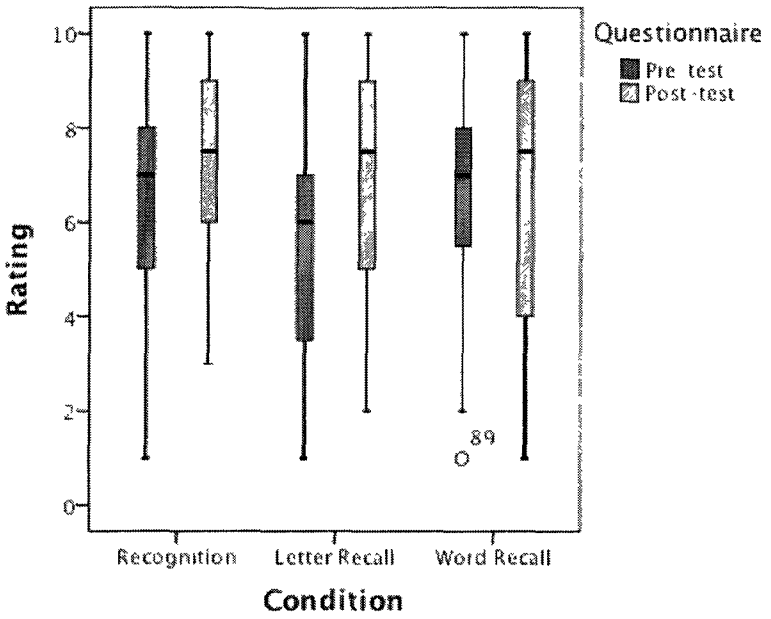


Figure 21: "Easy to remember" ratings per condition, before and after participation.

Participants were asked to rate how easy it was to remember their passwords for each authentication type on a scale from one to ten, where a score of one meant very difficult, and ten meant very easy. The recognition condition had the highest mean rating in both the pre- and post-task questionnaires, and witnessed the highest mean improvement. However, as Figure 21 helps explain, the difference among conditions was not found to be significant in either case, when tested using Friedman's test ($\chi 2 = 1.000$, $p = 0.607$ before; . $\chi 2 = 1.254$, $p = 0.534$ after).

Along the same scale, another three questions asked participants to rate the statement that each of the password types were easy to learn. The word recall condition was rated most highly (mean 6.97), followed by the recognition based passwords (6.60) and then the letter recall password type (6.31), These the observations for this question are visible in Figure 22. A Kruskal-Wallis test revealed that there was no significant difference among any of the means ($\chi 2 = 1.858$, $p = 0.395$).

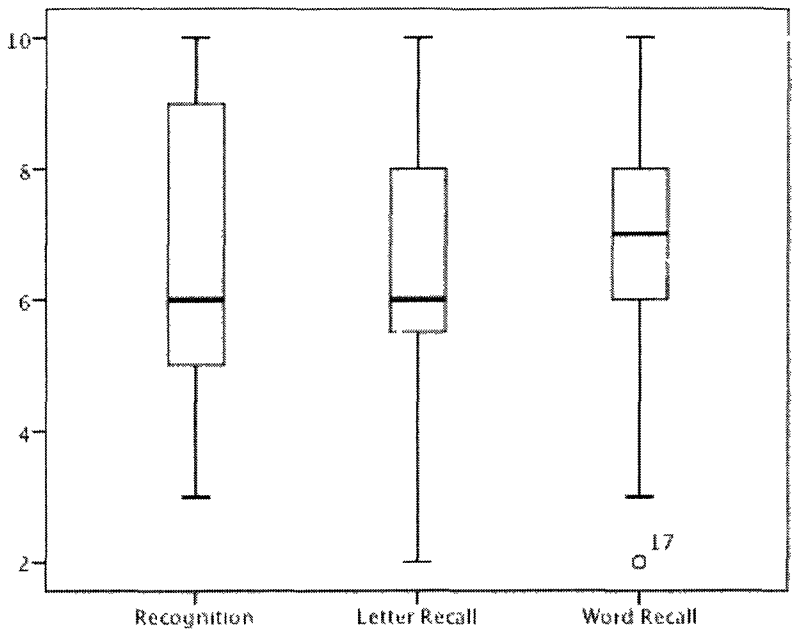


Figure 22: Boxplots rating how easy it was to learn passwords from each condition.

And finally, one question on the second survey asked participants which password system they preferred. Their selections appear in Figure 23:

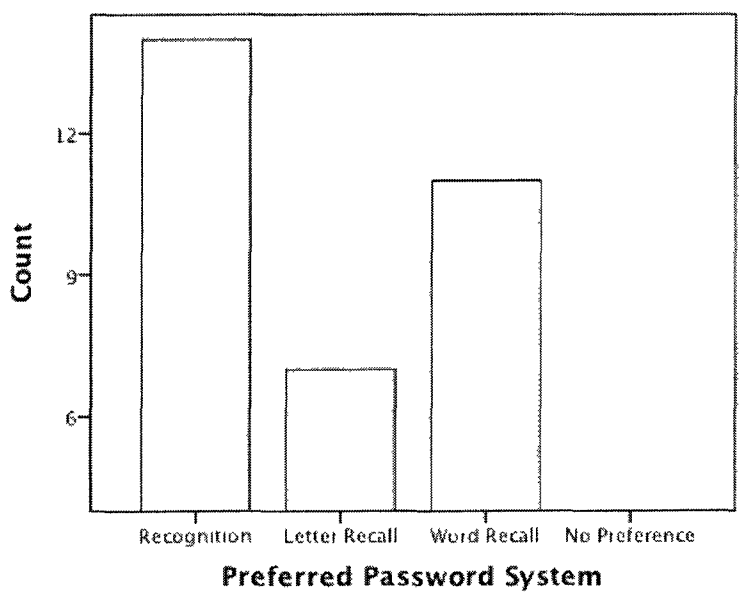


Figure 23: Frequencies of votes for preference of each authentication condition.

The graphed responses suggest a preference for the recognition condition, followed by the word recall and letter recall conditions. However, investigation via a chi-squared analysis showed no significant difference among the participants' choices ($\chi 2 = 6.444$, $p = 0.092$).

Experiment 2

Interference between passwords may play a role in the ability of participants to successfully retrieve their passwords from memory and submit them to the intended web sites in order to authenticate. Because interference may represent an important aspect of a password system's usability, a secondary investigation was conducted. This study involved assigning each participant three of the same type of password (either letter recall or recognition types), for use in the same three websites as those in the main study, and monitoring their authentication attempts for potential instances of interference. Our focus was on the potential of the novel recognition condition and that of the letter recall condition, which is most similar to present-day passwords, and word recall type passwords were omitted from this analysis.

This investigation used a between-subjects design consisting of two experimental conditions. Participants in one condition were assigned three letter recall type passwords, and participants in the other were assigned three recognition type passwords. Both conditions were used in the same three websites that were employed in experiment 1. Assigning participants three of the same types of passwords allowed us to observe evidence of password interference across the participants' three web sites.

The authentication condition again served as the independent variable (IV) in this evaluation. To investigate our hypothesis we needed to review the access logs of each web site and count the instances of interference, which served as the dependent variable (DV). For the purposes of this study we defined an instance of password interference as an event where at least

half of the contents of a password for one website are submitted as credentials in another (including passwords that may have been reset, and were previously valid). Instances of password interference were counted on each of the three websites, and mean interference was calculated for each participant.

Research Hypothesis

Recognition judgments have been described as more effective over lengthy periods of time than those based solely on recall. The recognition condition also presents users with *cues* to the content of their passwords, and limits the users' password selections to a small subset of potential words. It is believed that the word recognition condition would result in significantly fewer instances of password interference than the letter recall condition, which relies on the process of pure recall, as stated in hypothesis one:

$H1_1$: There will be significantly fewer instances of password interference witnessed in the word recognition authentication system when compared to the letter-recall based authentication mechanism.

$H1_0$: There will be no significant difference in the frequency of instances of password interference when using the recognition system vs. the letter-recall system.

Method

The investigation for this iteration of our study was conducted very similarly to the manner in which experiment 1 was carried out, albeit with a few key differences that are highlighted in the paragraphs that follow.

Participants.

Participants for this experiment were recruited in the same manner as those for the initial experiment, although it turned out that none were enlisted via Carleton's SONA system for online registration. As in experiment 1, we recruited people without significant visual impairment or other conditions, which could have impeded performance in this study.

Materials.

To administer this experiment we made use of the same three websites as in the earlier investigation (screenshots available in Appendix H). Automated reminders were sent to our participants at the same intervals of two and four days following recruitment (example in Appendix G), which asked them to visit each of the three websites and authenticate using the passwords they were assigned.

Apparatus.

This experiment made use of the same apparatus as was used and outlined for experiment 1. Participants used personal computers with Microsoft Windows and Internet Explorer when in the lab to learn their passwords and fill out the questionnaires. When browsing from outside the lab setting, participants were free to use any computer with an Internet connection. Password assignment, training and logs were managed via the MVP password framework (Chiasson et al., 2010).

Procedure.

Phase 1.

In this experiment, the procedure was largely identical to that of experiment 1. Participants arrived at the lab, were given an explanation of the study, an informed consent form,

and a participant information survey, which asked for demographic related data. The participants were then assigned their passwords, shown to the sites and given opportunity to practice. The only difference in this phase was that participants in this study were not asked to fill out a pre-test questionnaire.

Phase 2.

As in experiment 1, participants were asked to log into the sites between lab sessions and were sent two notification e-mails to this end at intervals of two and four days from their initial lab session. The web sites kept logs of all authentication related activity during this time, as well as the two lab sessions, for later analysis.

Phase 3.

Upon arriving at the lab, participants were greeted and asked to log into each site one last time. If they failed to do so, they were asked to make at least two attempts, without resetting their passwords. Participants were then debriefed and compensated, but were not asked to complete a post-task questionnaire.

Analysis Plan

While the logs for the three web sites collected all the same authentication related data as in experiment 1, we were exclusively interested in the interference related data. In order to count the interference events, every failed login attempt was contrasted with all passwords that belonged to the participant in question. For the recognition condition, a failed attempt that included at least three words from another password constituted interference, and for the letter recall type passwords a failed attempt that included at least 3 consecutive letters from another password was deemed a result of interference. A mean interference score was then calculated from the interference frequencies of each of a participant's three sites.

Hypothesis One.

$H1_1$: There will be significantly fewer instances of password interference witnessed in the word recognition authentication system when compared to the letter-recall based authentication mechanism.

$H1_0$: There will be no significant difference in the frequency of instances of password interference when using the recognition system vs. the letter-recall system.

To test this hypothesis, an independent samples t-test was conducted, to compare the mean instances of interference encountered in the recognition condition with those of the letter-recall condition. Authentication mechanism served as the IV, and the mean interference was the DV. In the event that the distributions did not satisfy the conditions of normality, a Mann-Whitney U test was used.

Results

Participants.

We recruited 20 participants for the investigation on password interference, and all of them completed the first session. In this secondary investigation of password interference, participants 1 and 20 were omitted due to corrupt data, and participant 6 was eliminated for a complete lack of participation. Hence eight people were assigned to the letter recall condition, and nine to the word recognition condition. No additional participants were recruited for this study.

Among the participants in this investigation, the average age was slightly older than 25 years, ten of whom were male and seven female. Eleven of them reported social science related

backgrounds, while six declared natural science or engineering related fields of study, and there was an average of 3.88 years of post secondary studies in this sample. Twelve people spoke English as a first language, and for five it was a second language. When asked to rate their computer skills on a scale from 1 to 10 where 1 meant "novice" and 10 meant "expert", the average response was 7.53, and 8 was the median answer; Nobody rated their skills less than 5, and all of them reported using the Internet daily.

Comparison with Experiment 1.

Before proceeding to the analysis of interference, it is worthwhile to briefly compare the results obtained in experiments one and two. Differences in the upcoming comparisons may be exacerbated by the different sample sizes and experimental designs associated with each experiment. Experiment one was conducted with a sample of 36 participants, while experiment two was conducted with the participation of 17 individuals. Experiment one was implemented using a within-subjects design in which each participant was assigned three passwords, each belonging to a different authentication condition, and experiment two necessitated a between-subjects design in which each participant was assigned three passwords, all of the same authentication type.

In order to compare the results of this experiment with those of experiment 1, an overall impression of the results for this experiment should first be considered. Descriptive statistics on the results from experiment two, divided by authentication condition have been calculated and can be seen in Tables 8 and 9 below.

Table 8: Descriptive Statistics for Letter Recall in Experiment 2.

	N	Min.	Max.	Mean	Median	Std. Dev.	Skewness Statistic	Skewness Std. Error	Kurtosis Statistic	Kurtosis Std. Error
Success	8	1	3	2.25	2.5	0.850	-0.389	0.752	-1.918	1.481
Failure	8	0	6.33	1.83	1	2.145	1.544	0.752	2.199	1.481
Resets	8	0	0.33	0.041	0	0.117	2.828	0.752	8	1.481
Login Time	8	5.67	17.33	9.58	8.385	3.943	1.166	0.752	0.941	1.481
Mem. Time	8	49.27	200.89	140.64	153.51	53.616	-0.579	0.752	-0.714	1.481
Interference	8	0	5	1.13	0.33	1.799	1.852	0.752	2.87	1.481

Table 9: Descriptive Statistics for Recognition in Experiment 2.

	N	Min.	Max.	Mean	Median	Std. Dev.	Skewness Statistic	Skewness Std. Error	Kurtosis Statistic	Kurtosis Std. Error
Success	9	1	4.33	2.851	3	1.001	-0.374	0.717	0.459	1.4
Failure	9	0	7	1.667	0.67	2.266	1.966	0.717	3.93	1.4
Resets	9	0	1	0.222	0	0.441	1.62	0.717	0.735	1.4
Login Time	9	28.11	277.44	67.079	37.08	80.156	2.832	0.717	8.187	1.4
Mem. Time	9	61.8	198.3	147.646	146.02	49.056	-0.846	0.717	-0.335	1.4
Interference	9	0	1.66	0.221	0	0.551	2.798	0.717	7.979	1.4

The results from experiment two will be shown with respect to the dependent variables used in the hypotheses considered in experiment 1, and plotted alongside the corresponding results from that experiment. Hypothesis one was related to the duration of time for which each password was successfully remembered. Figure 24 depicts the distributions of memory time within their respective conditions.

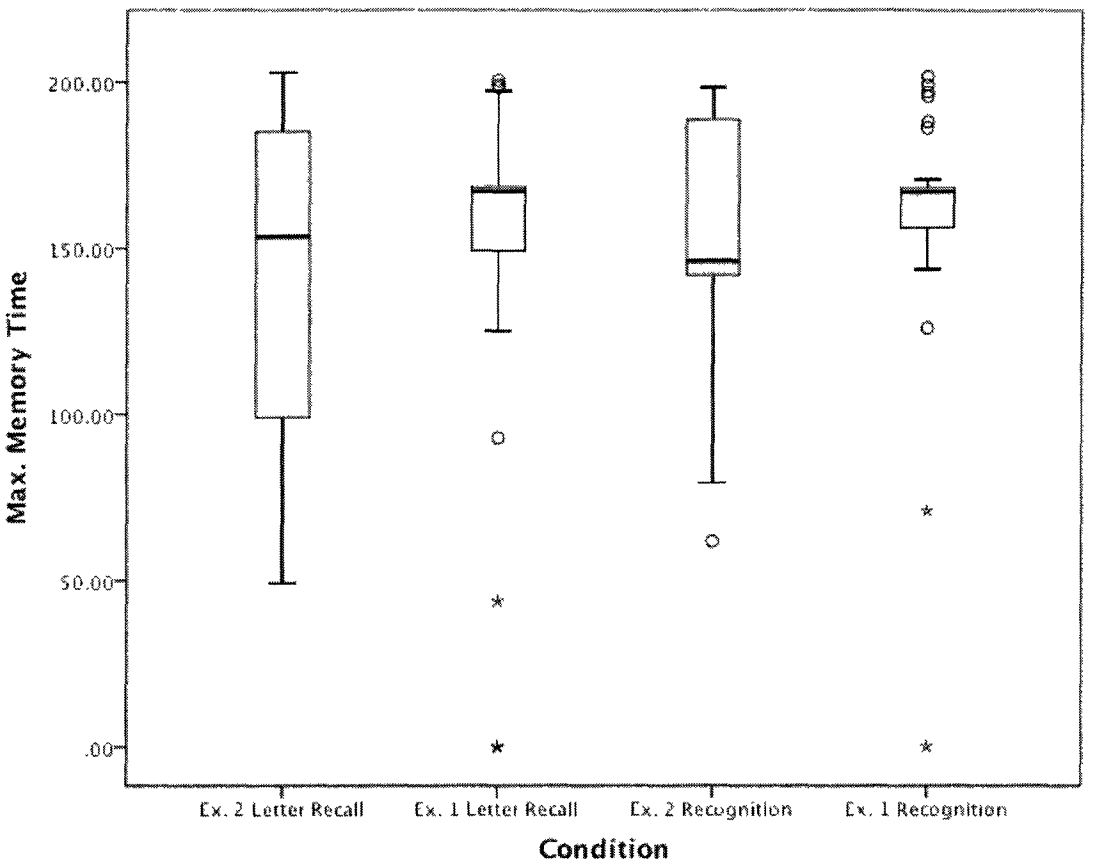


Figure 24: Boxplots of maximum memory time by condition, per experiment.

Given the differences between these studies, the distributions of memory time appear reasonably comparable. Hypothesis two focused on the number of password resets initiated in each authentication condition. Figure 25 displays those observations.

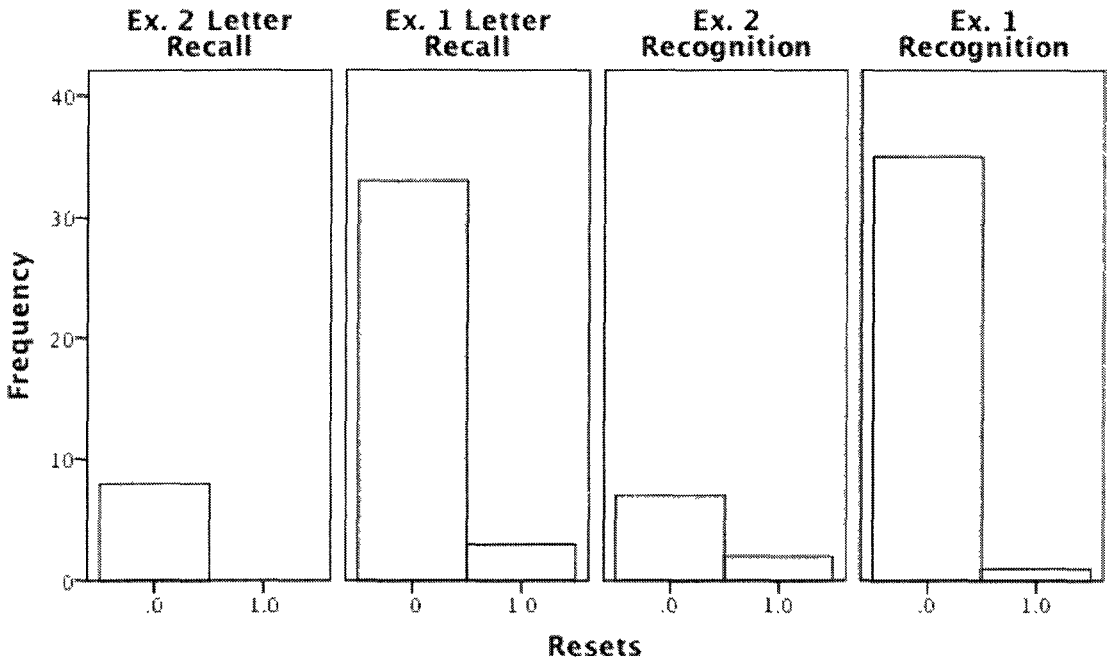


Figure 25: Histograms of password resets performed per condition, in each experiment.

Even though there appears to be a vast difference between the conditions in the case of password resets, notice that participants reset their passwords once at most, and only a few people reset their passwords, even in the poorest case. Of course, experiment 1 had many more participants than experiment 2. Hypothesis four is the final memorability related evaluation. In experiment 1, an evaluation was conducted in order to assess a difference in the number of passwords in each condition that were remembered for the duration of the study. Figure 26 permits a comparison between the observations in that investigation, with those of experiment 2.

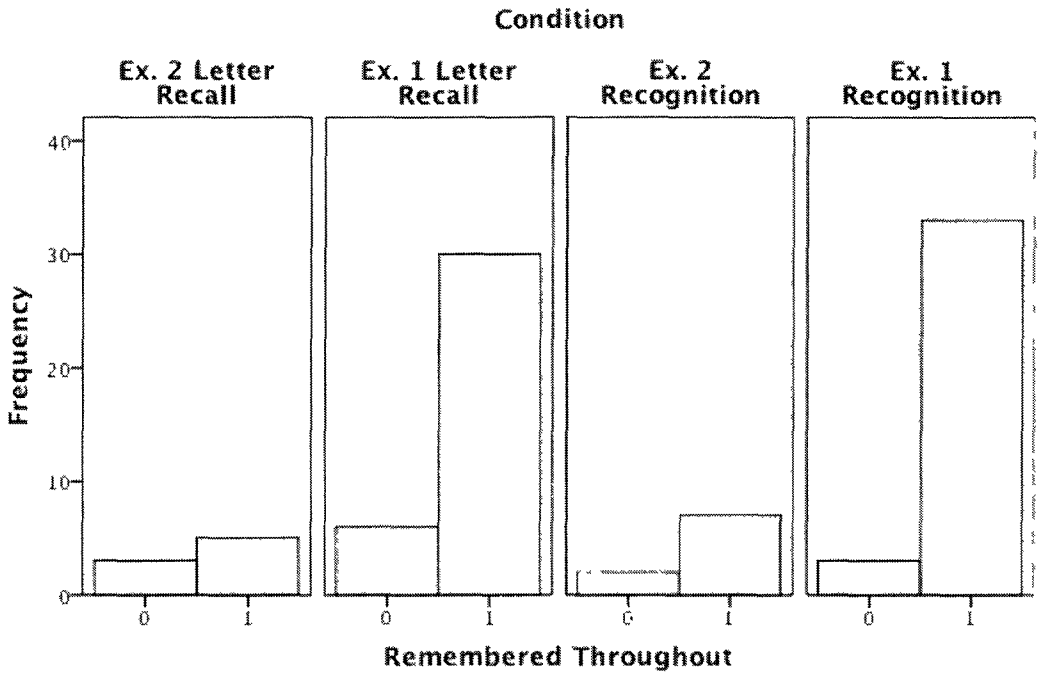


Figure 26: Histograms of passwords remembered for the duration of the study, by condition.

The third hypothesis in experiment one was related to performance. In this evaluation, the time it took participants to successfully log in was compared across conditions. The same comparison can now be made, including the observations from experiment 2. Figure 27 permits a comparison of the conditions.

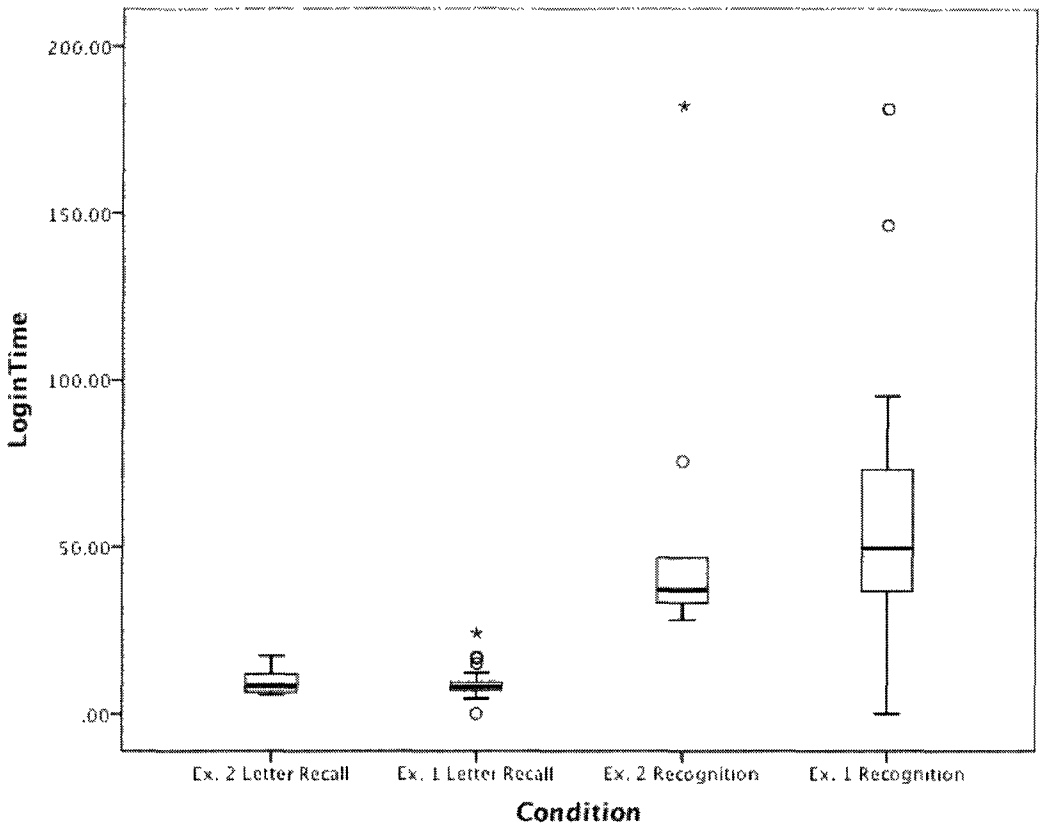


Figure 27: Boxplots of login times observed by condition, per experiment.

Despite the great differences between the two experiments, the results show that the time required to login is very different between the letter recall and recognition conditions.

Having considered the differences between the designs of the experiments and reviewing the observations common to both, we now proceed to the evaluation of the hypothesis for experiment 2, which examines the potential for differences in password interference between the recognition and letter recall conditions.

Hypothesis One.

Upon completion of experiment 1, it became clear that the resulting observations did not produce a normal distribution, as seen in Figure 28. This violation of the assumption of normality dictated that we investigate any significant difference among the groups using a Mann-Whitney U test.

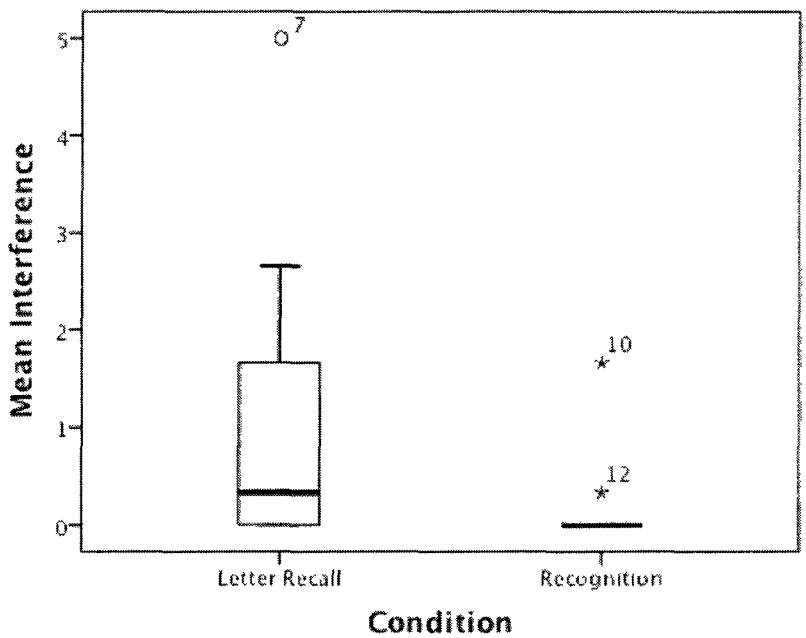


Figure 28: Boxplots of mean password interference

While Figure 28 appeared to suggest there was a difference in means between the two authentication conditions in terms of password interference, it was not found to be significant ($U = 20.500$, $p = 0.094$). This seems surprising because the recognition-based authentication condition is in fact more resistant to interference by design. In all of the password schemes in this study, passwords are assigned and must be entered in the correct order. In either of the recall mechanisms participants are free to enter whatever text they choose into any of the text boxes displayed to them, however, in the recognition condition participants are forced to choose one word at a time from a set of 26 shown per panel, making the possibility for complete password

interference extremely unlikely. This experiment showed no evidence for reduced password interference in the recognition condition, however this issue may be deserve further study.

Discussion

The three password systems used in this study each made use of text-only passwords. Passwords of the first type were composed of six randomly generated letters that were to be typed in succession in order to authenticate. Among the three conditions used in this study, this one most closely approximated passwords in use today, and involves a pure recall situation. The second password type consisted of four randomly assigned whole words, which the user had to type out to submit. This also presented the user with a pure recall situation, but with greater potential for semantic meaning in their content. The third password scheme was a cognometric password mechanism, except that where graphical images would normally be used in sets of tiles displayed in several panels, pictures of words were displayed instead. This preserved the recognition aspect of a cognometric password system, while eschewing the potentially confounding use of images.

This study was able to focus on the usability of the password mechanisms themselves because the security level was held constant across the three conditions. That is to say that in each case, the complete set of potential passwords was the same size. The password space was 28 bits in strength, which is acceptable in some text-based password systems in the real world. Recently, passwords of just 21-bit strength were suggested to be sufficient for protection against the majority of attacks online (Florencio & Herley, 2010). Many text password systems in use today have theoretical password spaces that are much larger than the space allotted in this study, however, they allow for user chosen passwords, and these same users will often create passwords

that only meet a system's minimum requirements, resulting in passwords of comparable strength to the security level used in this study or weaker.

Considering memorability, the recognition and letter recall conditions produced similar results, and both of them typically outperformed the word recall condition. In the end, there was no significant difference in maximum memory time between the three conditions, there was a significant difference between the recognition and the word recall conditions in terms of password resets, and there was also a significant difference between recognition and word recall when considering the number of passwords remembered for the duration of the study. The notable skew in the distribution of the word recall condition indicates that while many were able to remember that type of password for a considerable period of time, many people had a great deal of trouble remembering that password for a meaningful duration, and this stood in contrast to the distributions associated with the other two conditions for the test of that hypothesis.

We had anticipated that the recognition condition would prove significantly more memorable than the two recall based conditions, but that difference did not materialize. In the test of maximum memory time, a ceiling effect was witnessed among the distributions. Quite a few participants remembered their passwords for more than six days. It is possible that a longer duration of participation time in this study would have allowed for greater differentiation between the conditions and results more closely resembling a normal distribution, but this is unclear. Normally distributed results would have facilitated the use of statistically stronger tests, however, a greater period of time between lab sessions may not have differentiated the letter recall from the recognition condition, because they performed so similarly in this experiment.

Perhaps the most surprising finding from among the memorability results is just how poorly participants fared in the word recall authentication condition. The letter recall condition served as a pure recall form of authentication and the recognition condition allowed a

comparison of recognition and pure recall. The word recall scheme, also a pure recall scheme, was included in the study because it potentially represented an enhanced form of recall over the letter recall condition. Requiring participants to recall whole words as a password, we anticipated would allow for increased semantic meaning of the password content and thus more thorough encoding process, but the significantly higher mean number of resets and fewer number of passwords remembered for the duration of the study highlights the fact that this was not the case. This condition presented an additional source of error in the misspelling of words; however, we observed that spelling mistakes were not a common source of error.

It was surprising that recognition as retrieval mechanism showed no greater memorability than the letter recall scheme. The two factors that bolster the idea of using cognometric graphical passwords as a successor to text-based passwords were the pictorial-superiority effect, and recognition rather than recall. Our study eliminated the possible confound by omitting the use of pictures from the experimental conditions. The revelation that recognition did not enhance the memorability of text-based passwords suggests that perhaps it is the use of pictures alone that may improve the usability of password systems.

Another possible explanation rests with the distinctiveness of the words chosen. A similar study was conducted last year (Hlywa, 2010) comparing the usability of different kinds of images in cognometric graphical password schemes. In that research, the distinctiveness of each image in the password set was suggested to be the principle factor in the learning and retrieval of passwords. Building upon this line of thought, perhaps the composition of the words in our word set could be made more distinctive, for example in terms of semantic meaning and shape (word length and spelling), and also by using bold, italics or styling in random cases. This concept of distinctiveness implies that unique perceptual properties enhance memorability (Hunt & Elliot, 1980). Furthermore, in the context of words, a distinctive orthographic to phonologic mapping

has also been demonstrated to enhance memorability (Hirschman, Elliot, Jackson & Eric, 1997). This issue may have influenced the word recall condition as well as the recognition condition.

The time required to login successfully revealed the most glaring difference between the conditions. The recognition condition required significantly longer for its users to login than that of either recall based condition. The time to login in the word recall condition was roughly twice that of the letter recall condition, which seems appropriate, given that participants had to type four times as many characters, however the time to submit a recognition-based password was more than double that of word recall. When the sizeable differences in performance times across conditions was revealed, the login times were investigated using only data from the second lab session. This would allow for the maximum amount of repetition and less chance for distraction, but these same results were mirrored in that analysis.

The influence of distraction available in this scenario is one source of concern. Participants may have seen a word with a unique meaning to them and begun to think about that for a moment. Alternatively, the volume of words in front of them may have been too great, causing the participant to attempt to recall their password instead, and then search laboriously for the word they had recalled.

The great length of time required to login in the recognition condition seems to suggest that although recognition decisions can in general be made more quickly than attempts at recall, the user is making a sufficient number of recognition decisions to make the process slower than pure recall, hindering the system's usability.

Reducing the number of distractor words per panel, and increasing the number of panels per password is worthy of some attention in order to maximize the speed per panel, while preserving a desirable level of security. Presenting two words per panel would result in extremely fast processing of the panels, but may require processing a prohibitive number of

panels in order to authenticate. The PassFaces scheme, involving 9 words per panel and 4 panels, appears very usable but with an associated poor level of security. Having 26 words per panel in our recognition condition drove some participants to attempt to recall their words rather than relying on recognition and incited a serial search through the words to find the words they may have been able to remember. There could be an ideal combination of words per panel and number of panels, which might allow for a quick login at an acceptable level of security.

Recall that the words used in our recognition condition were shown in different positions every time an authentication attempt was made. The same words were always used on the same panels, but those words were always shuffled. Assigning passwords maximized entropy in order to combat guessing or brute-force attacks, and shuffling is implemented in cognometric password schemes as a preventative measure to counter capture-based attacks such as shoulder surfing. In this investigation, shuffling also ensured a pure recognition scenario. However, if we were less concerned with shoulder surfing or more sophisticated capture attacks, we could do away with the shuffling of the paneled words.

This shuffling may indeed have been the element that caused the recognition condition to demand additional effort, resulting in its burdensome login times. Eliminating the shuffling may have enhanced the memorability of the passwords, allowing users to recognize the whole panel of words, and first *recall* the general position of their words, and then allowing recognition of the sought after words themselves. Over time people could have even developed a spatial memory regarding the positions of their words, or muscle memory for the motion of their hands positioning the cursor. This may have improved upon the terribly poor login times associated with this condition. The authentication condition would no longer be an example of pure recognition, but this would not be of concern to the user. The added ability to recall information

about the panels would combine the two retrieval mechanisms and possibly result in the creation of a more usable authentication mechanism.

In the context of passwords, memorability is a large part of their usability. As the number of passwords we use grows, individual passwords may become less memorable. Novel password schemes may present themselves as more memorable because of an artificial lack of any similar passwords in memory. The mistaken submission of a valid password from one application into another system would result in an unsuccessful login attempt. This is referred to as password interference, and is a failure to retrieve the correct password from memory.

A great deal of research has been conducted on the role of interference in memory, and some recent work in the specific context of authentication (Chiasson, et al., 2009). While text based passwords have existed for decades, the present study is the first to substitute words into a cognometric password mechanism. Since interference can severely hamper the usability of text based password schemes it is important to investigate the potential for interference in the proposed recognition-based password system.

Experiment 2 was conducted specifically to address the possibility of password interference in the novel recognition condition by comparing interference observations in that condition to observations in the letter recall condition. The data revealed no significant difference between conditions. However, The low number of participants in this study may have been a key limitation, and this result may merit further investigation. Participant observations indicated there were a greater number of occurrences of interference in the letter recall condition, however, complete failure to remember the appropriate words or letters was the bigger issue for either condition. When learning their passwords, upon realizing that a word or two in their new password was also used in one of their other passwords, some seemed relieved, as though it would simplify the learning process, and others immediately realized that this would add to the

difficulty. In the end, these duplicate words only seemed to complicate things, adding further distraction in the recognition condition (participants would pause to point out the re-used word), and complicating the word recall condition by causing them to consider whether or not any of their other words were substituted in by mistake.

An interference evaluation of the word recall condition was not done, which would have been interesting, however given these findings it is unlikely there would have been a significant difference there.

When considering participant behaviour, it appeared that participants exerted comparable amounts of effort in each authentication condition. There were no significant differences in login attempts overall, or successful attempts across conditions. Practice didn't make perfect in this study either, as the number of login attempts per successful authentication was not significantly different between conditions, or over time. Upon concluding their involvement in the study, participants were asked if they had written their passwords down. Four of them acknowledged that they had, although it seemed that the majority had only done so as a form of practice during the learning phase. Additionally, it appeared that writing down passwords was more commonly used as a strategy for this in the recognition and word recall conditions, which used whole words.

Individual differences among the participants were controlled for in the within-subjects design of the first experiment. We had anticipated that the recognition condition would fare best in the memorability related tests. However, we acknowledge that while some people may prefer a recognition scenario, others may perform best in a situation involving recall. Since the recognition scheme did not improve usability uniformly, perhaps the process of authenticating could be improved by accommodating the preferences of each user. This might possibly be

accomplished by allowing users to choose which type of authentication they would like to use on each site or service they use.

Participants were observed using several strategies to aid in the memorization and retrieval of their passwords. In the letter recall condition, participants tried to pronounce their random letter passwords as whole words, and commented that these passwords were easiest to remember when there were vowels present, making them easier to sound-out. Jung (1968) found that meaningful syllables were easier to remember than non-meaningful syllables. Participants were thus attempting to ascribe meaning to their meaningless random letter passwords.

People like to use whole words in their passwords, and so it was expected that the word recall authentication mechanism would produce better results than it did. Using whole words allows people to group the letters composing their passwords and capitalize on the phenomenon known as memory chunking (Miller,1956).While chunking is put to use to group the letters of a known word, participants must not have been able to treat their sets of four words as groups or chunks of words. For example they may have created sentences composed of their four words. The significant difference in password resets, however, leads us to believe that chunking did not help participants learn or remember their word recall based passwords. Instead, it was observed that some people tried to remember the first letters of each of their words, which offered little help in either the word recall or recognition scenarios. Indeed, many of the mistakes made included words with the same first letter.

In the word recall and recognition conditions some participants tried to compose sentences or stories from their password set. Participants commented that when their recognition password sets included verbs they were easier to remember, and likewise when words "belonged together" (The words "tongue" and "throat" were given as one example). So, similar semantic meaning of words therefore played a role in the memorability of these passwords, which is

consistent with previous findings (Deese, 1959). Interestingly, the ratio of verb words to the set of all words in the selection set was much lower than the ratio of vowels to the size of the alphabet, which may have made the whole word passwords more difficult to group. It therefore may be possible to better facilitate this strategy for password memorization.

Conclusion

The Internet has become an integral part of our lives, both personal and professional. As the number of web sites and services continue to grow, so does the ubiquity of text-based password systems. Text-based passwords can be difficult to remember and use, especially given the increasing number of passwords we are expected to initialize and use regularly, and the wide variety of security policies to which we must adhere.

Cognometric graphical password systems have been lauded because they may capitalize on both the pictorial superiority effect, and recognition as a retrieval process, which is regarded as superior to recall. While the former cannot be applied to text-based passwords, we sought out to discover whether or not recognition alone could enhance the usability of text-based password mechanisms.

This study sought to determine if the use of recognition in text-based authentication systems could improve their usability. The experiment involved assigning three different types of passwords to participants to use on three websites that we could monitor, for a period of one week. One form of password consisted of 6 randomly generated lower case letters, and another type consisted of four randomly generated whole words. Both of these mechanisms use recall as retrieval mechanism, with the difference being that the whole word condition would involve a great deal more semantic information, ideally simplifying retrieval of the password. The third password mechanism closely resembled a cognometric graphical password system with 26 images per panel, except that in our case the images were simply pictures of words. With the password space held constant across conditions, we were then able to compare the system based on several usability metrics, and through our participants' feedback.

No significant differences were observed in maximum memory time across conditions. The recognition condition produced significantly fewer password resets than the word recall condition, as did the letter recall condition. In terms of the number of passwords that were remembered for the duration of the study, the recognition condition performed comparably to the letter recall condition, and significantly outperformed the word recall condition, though the letter recall condition did not. The surprising weakness of the recognition condition was the time required to login. All differences were significant in this test, where the letter recall condition performed best, followed by the word recall condition and then the recognition condition. Even if implementing recognition resulted in more memorable passwords, in this present form it would also make them more time-consuming to submit.

We also conducted a secondary study to investigate the potential problem of password interference in this recognition-based text password mechanism. Participants were either assigned three letter-recall type passwords, which most closely resemble passwords in use today, or three of the recognition-based passwords. No significant difference in the occurrence of password interference was observed.

The impetus for this study was the confound created in the comparison of the usability of cognometric graphical password systems to that of traditional text-based password systems. Previous studies which incorporated this confound have demonstrated increased usability in graphical password mechanisms relative to text-based authentication. In light of this, and considering that we failed to support the influence of recognition rather than recall as retrieval mechanism as the factor enhancing usability, perhaps it is the use of pictures rather than text that potentially renders graphical password mechanisms more effective than the traditional text password system. Further study of this issue is necessary.

An alternative interpretation may relate to the distinctiveness of words comprising a password set, as distinctiveness of words may lead to enhanced memorability. The shuffling of the paneled words in our recognition condition ensured that this was a pure recognition scheme, however it seemed to add undue difficulty to the mechanism. A scheme allowing users to capitalize on all manners of memory retrieval may in fact represent an optimally usable authentication mechanism.

This study was subject to two obvious limitations. While we strove for ecological validity in the design of this experiment, the period of time for which we could monitor password use was limited to one week, and the motivation to remember the passwords or reset them when forgotten was not critical. For greater ecological validity, a longer term study would be desirable. The first experiment involved 36 participants and the second involved 17. Sample size is thus a potential factor in the lack of significance in some of the hypothesis tests, and this limitation could be addressed by conducting studies with greater numbers of participants in future.

Having completed this study, there are a few investigations that could naturally follow. The issues of study duration and sample size have been addressed. A direct comparison between the recognition condition used in this study with one using a graphical cognometric scheme of the same password space, to contrast the use of words with pictures would also be very interesting. This study's recognition condition suffered from terribly long login times. Some investigation adjusting the number of images per panel, and the number of panels per password seems like another investigation of our cognitive abilities that would be appropriate for password usability. Finally, in order to create a satisfactory interface for users, recognition might be implemented in such a way that it would not hinder people from logging in as fast as they can

with their current passwords. It may be that an approach allowing for both recognition and recall would be ideal.

References

Adams, A. & Sasse, M. A. (1999). Users are not the enemy. *Communications of the ACM*, 42(12), 41 – 46.

Anderson, J. R., & Bower, G. H. (1972). Recognition and Retrieval Processes in Free Recall. *Psychological Review*, 79(2), 97 – 123

Anderson, J. R., & Bower, G. H. (1974). A Propositional Theory of Recognition Memory. *Memory & Cognition*, 2(3), 406 – 412.

Bauer, J. L. (2008, October). Ogden's Basic English. Retrieved November, 2010, from http://ogden.basic-english.org/

Bernbach, H. A., (1967). Decision Processes in Memory. *Psychological Review*, 74(6), 462 – 480.

Biddle, R., Chiasson, S., & van Oorschot, P. C. (in press). Graphical Passwords: Learning from the First Twelve Years. *ACM Computing Surveys*.

Brostoff, S. & Sasse, M. A. (2000). Are PassFaces More Usable Than Passwords? A Field Trial Investigation. *British Human-Computer Interaction Conference (HCI)*, September 2000.

Burr, W. E., Dodson, D. F., Polk, W. T., Evans, D. L. (2004). Electronic Authentication Guideline, in NIST Special Publication 800-63.

Chiasson, S., Biddle, R. & van Oorschot, P. C. (2007). A Second Look at the Usability of Click-Based Graphical Passwords. *Symposium on Usable Privacy and Security (SOUPS),* Pittsburgh, PA, U.S.A.

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., van Oorschot, P. C. & Biddle, R. (2009). Graphical Passwords: Learning from the First Generation. Technical Report TR-09-09, School of Computer Science, Carleton University, Ottawa, Canada.

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.

Craik, F.I.M. & Lockhart, R.S. (1972). Levels of processing. A framework for memory research. Journal of Verbal Learning and Verbal Behaviour (11), 671 – 684.

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

Crowder, R. G. (1976). *Principles of Learning and Memory*. New Jersey: Lawrence Erlbaum Associates.

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), 128 – 152.

Deese, J. (1959). Influence of inter-item associative strength upon immediate free recall. Psychological Reports (5), 305-312.

Deese, J. (1961). From the isolated verbal unit to connected discourse. In C.N. Coffer (Ed.), *Verbal learning and Verbal Behavior* (pp. 11-31). New York: McGraw-Hill.

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

Dunphy, P. & Yan, J. (2007). Do Background Images improve "Draw a Secret" Graphical Passwords? *Proceedings of the ACM conference on computer and communications security*. pp. 36 – 47.

Eagle, M. & Leiter, E. (1964). Recall and recognition in intentional and incidental learning. *Journal of Experimental Psychology*. 68, 58 – 63.

Ebbinghaus, H. (1885). Memory: A Contribution to Experimental Psychology. Classics in the History of Psychology. http://psy.ed.asu.edu/~classics/Ebbinghaus/index.htm. Accessed January, 2011. Site created by Christopher D. Green, Toronto, ON.

Estes, W.K. & DaPolito, F. (1967). Independent variation of information storage and retrieval processes in paired-associate learning. *Journal of Experimental Psychology*. 75, 18 – 26.

Ferguson, N. & Schneier, B. (2000). A Cryptographic Evaluation of IPSec. Counterpane Internet Security, Inc. San Jose, California.

Federal Information Processing Standards Publication (FIPS) (1985). Federal Information Processing Standards Publication 112: Password Usage, *National Institute of Standards and Technology,* http://www.itl.nist.gov/fipspubs/fip112.htm. Accessed Jan, 2011.

Florencio, D. & Herley, C. (2010). Where do security policies come from? *Proceedings of the Sixth Symposium on Usable Privacy and Security*. ACM: Washington.

Gardiner, J.M. (1988). Functional aspects of recollective experience. *Memory and Cognition.* 16(4), 309 – 313.

Graf, P. & Schacter, D.L. (1985). Implicit and Explicit Memory for New Associations in Normal and Amnesic Subjects. *Journal of Experimental Psychology: Learning, Memory and Cognition.* 11(3), 501 - 518.

Herley, C., van Oorschot, P.C. & Patrick, A.S. (2009) Passwords: If we're so smart, why are we still using them? In R. Dingledine and P. Golle (Eds.): FC 2009, LNCS 5628, pp. 230–237, 2009. Microsoft Research, Redmond, U.S.A; School of Computer Science, Carleton University, Canada; National Research Council, Ottawa, Canada.

Hintzman, D.L. (1990) Human learning and memory: Connections and Dissociations. *Annual Review of Psychology*. 41, 109–139.

Hirshman, E., & Jackson, E. (1997). Distinctive perceptual processing and memory. *Journal of Memory and Language, 36*(1), 2-12.

Hollingworth, H.C. (1913) Characteristic differences between recall and recognition. *American Journal of Psychology*. 24, 532 – 544.

Hunt, R. R, & Elliot, J. M. (1980). The role of nonsemantic information in memory: Orthographic distinctiveness effects on retention. *Journal of Experimental Psychology: General, 109*(1), 49-49-74.

Hlywa, M.A.X., (2010). Do houses have faces? The effect of image type in recognition-based graphical passwords. Carleton University Department of Psychology, 2010.

Jacoby, L. L. (1983). Perceptual enhancement: Persistent effects of an experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9, 21-3.

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, Washington, D.C.

Johnston, W.A., Dark, V.J., & Jacoby, L.L. (1985). Perceptual fluency and recognition judgments. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 11(1), 3 – 11.

Jung, J. (1968). Verbal Learning. New York: Holt, Rinehart & Winston.

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

Luh, C. W. (1922). *The conditions of retention*. Psychological Monographs. 31(3), 1-89.

Miller, G.A. (1956), The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. Psychological Review, 63, 81-97.

Moncur, W. & LePlâtre, G. (2007). Pictures at the ATM: Exploring the usability of multiple graphical passwords. *Human Factors in Computing Systems (CHI)*. San Jose, California, USA.

Norman, D.A. (1968). *Toward a theory of memory and attention*. Psychological Review. 75, 522 – 536.

Passfaces Corporation, "The science behind PassFaces," White paper, http://www.passfaces.com/enterprise/resources/white_papers.htm, accessed December 2010.

Postman, L., & Rau, L. (1957). Retention as a function of the method of measurement. *University of California Publications in Psychology, Berkeley*. 8, 217 – 270.

Postman, L. (1964). Short-term memory and incidental learning. In A. W. Melton (Ed.), *Categories of human learning*. New York: Academic Press.

Renaud, K. (2009). Guidelines for Designing Graphical Authentication Interfaces. *International Journal of Computer Security (IJICS)*. 3(1), 60 – 85.

Renaud, K. & De Angeli, A. (2009). Visual Passwords: Cure all or snake oil? *Communications of the ACM*. 52(12), 135 – 140.

Richardson-Klavehn, A., & Bjork, R.A. (1988) Measures of memory. *Annual Review of Psychology*. 39, 475 – 543.

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

Sasse, M.A., Brostoff, S. & Weirich, D., (2001). Transforming the 'Weakest Link' – A Human/Computer Interaction Approach to Usable and Effective Security. BT Technology Journal.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning. Memory, and Cognition.13.* 501-518.

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

Suo, X. & Zhu, Y. (2005). Graphical Passwords: a survey. *Proceedings of the 21st Annual Computer Security Applications Conference*. pp. 463-472.

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

Tulving, E., & Pearlstone, Z. (1966) Availability vs. accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*. 5, 381 – 391.

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science, 247,* 301-396.

Tulving, E., & Thomson, D.M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*. 80, 352 – 373.

Tulving, E. (1975). Ecphoric processes in recall and recognition. In J. Brown (Ed.), *Recall and recognition*. London: Wiley.

van Oorschot, P. C., Vanstone, S. A. & Menezes, A. J. (1996). Handbook of Applied Cryptography. Boca Raton, FL, USA: CRC Press, Inc.

Watkins, M. & Gardiner, J. M. (1979). An appreciation of the generate-recognize theory of recall. *Journal of Verbal Learning and Verbal Behavior*, 18, 687–704.

Wickelgren, W.A., & Norman, D.A. (1966). Strength models and serial position in short-term recognition memory. *Journal of Mathematical Psychology*. 3, 316 – 347.

Wiedenbeck S., Waters, J., Birget, JC., 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.

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.

Yntema, D.B., & Trask, F.P. (1963) *Recall as a search process*. Journal of Verbal Learning and Verbal Behavior. Vol. 2, pp. 65 – 74.

Appendix A: Consent Form – Interactive Web Site Study (To be collected at the beginning of Session 1)

Research Personnel

Nicholas Wright Principle Investigator Carleton University (613) 520-2600 Ext. 6317 nwright@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 memorability of different forms of text passwords. The usability, security, and effectiveness of the two types will then be evaluated, compared and contrasted.

Task Requirements

At the outset of this experiment, we will assign to you a series of text passwords and ask that you remember them. We will ask you to login to some web sites we have created using your passwords, and fill out a brief questionnaire. During your second visit, in about one week, you will do the same thing. Between visits, we will ask that you login to the corresponding websites approximately twice per password.

Duration and Locale

The first session should take approximately 1 hour and the second approximately 30 minutes. You will be asked to log into, and make a short entry on, four websites throughout the time between appointments. You will be paid a $20 honorarium OR receive course credit for your time, upon completion of the second session. The testing will take place primarily at the HotSoft lab located in room 2110 HCI.

Potential Risk/Discomfort

There will be no psychological or physical risk from participating in this experiment.

Anonymity/Confidentiality

All data that is collected will be held completely confidential. The data will only be made available to those people involved with this testing. Data will be coded for identification purposes. All email addresses collected for the study will be discarded at the end of your second session.

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 $20 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. Monique Sénéchal Chair, Carleton University Ethics Committee for Psychological Research Carleton University (613) 520 2600 ext 1155 monique_senechal@carleton.ca	Dr. Janet Mantler Chair, Department of Psychology Carleton University (613) 520 2600 ext 4173 psychchair@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: _____ E-mail: _____

Participant's Signature: _____

Researcher's Name: _____

Researcher's Signature: _____

Date: _____

Appendix B: Participant Information – Interactive Web Site Study

(To be completed at the end of Session 1)

Welcome to the Interactive Web Site Study. Please tell us a little about your background and computer use in the questions that follow.
There are 9 questions in this survey

1 *Login* Name:

2 What is your age?

3 Gender:

- ○ Female
- ○ Male

4 Field of Study:

5 Number of years of post secondary education completed:

6 If currently studying, what year are you in?

7 What is your first language?

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

1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○

9 How often do you browse the web?

- ○ Daily
- ○ Several times per week
- ○ Once a week
- ○ Less than once a week

Appendix C : Interactive Web Site Study Post-Task (Session 1) Questionnaire

There are 10 questions in this survey

1 Login Name:

2 On a scale from 1 to 10, where 1 means you strongly disagree and 10 means you strongly agree, please rate the following statements:

	1	2	3	4	5	6	7	8	9	10
It will be easy to remember the clickable passwords	○	○	○	○	○	○	○	○	○	○
It will be easy to remember the random character passwords	○	○	○	○	○	○	○	○	○	○
It will be easy to remember the typed whole-word passwords	○	○	○	○	○	○	○	○	○	○

3 On a scale from 1 to 10, where 1 means you strongly disagree and 10 means you strongly agree, please rate the following statements:

	1	2	3	4	5	6	7	8	9	10
Being assigned my password is frustrating	○	○	○	○	○	○	○	○	○	○
I believe I will be able to remember all my passwords until the second appointment	○	○	○	○	○	○	○	○	○	○

4 On a scale from 1 to 10, where 1 means you strongly disagree and 10 means you strongly agree, please rate the following statements:

	1	2	3	4	5	6	7	8	9	10
I believe the clickable password system is secure	○	○	○	○	○	○	○	○	○	○
I believe the random character password system is secure	○	○	○	○	○	○	○	○	○	○
I believe the typed whole-word password system is secure	○	○	○	○	○	○	○	○	○	○

5 On a scale from 1 to 10, where 1 means you strongly disagree and 10 means you strongly agree, please rate the following statements:

	1	2	3	4	5	6	7	8	9	10
It was easy to learn the clickable passwords	○	○	○	○	○	○	○	○	○	○
It was easy to learn the random character passwords	○	○	○	○	○	○	○	○	○	○
It was easy to learn the typed whole-word passwords	○	○	○	○	○	○	○	○	○	○

6 Approximately how many web sites do you visit that require a username and password?

7 Do you sometimes re-use the same password on different web sites?

○ Yes
○ No

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

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

9 On a scale from 1 to 10, where 1 means you are not at all concerned and 10 means you are very concerned, please rate the following statement:

How concerned are you about the security of your passwords?	○	○	○	○	○	○	○	○	○	○

10 If you had to create a new password for your bank account using a text password system, how would you go about choosing a new password?

Appendix D: Interactive Web Site Study Post-Tasks (Session 2) Questionnaire

There are 7 questions in this survey

1 Login Name:

2 On a scale from 1 to 10, where 1 means you strongly disagree and 10 means you strongly agree, please rate the following statements:

	1	2	3	4	5	6	7	8	9	10
The clickable passwords were easy to remember	○	○	○	○	○	○	○	○	○	○
The random character passwords were easy to remember	○	○	○	○	○	○	○	○	○	○
The typed whole-word passwords were easy to remember	○	○	○	○	○	○	○	○	○	○

3 Which of the three types of passwords did you prefer? Please choose **only one** of the following:

- ○ Clickable passwords
- ○ Random character passwords
- ○ Typed whole-word passwords
- ○ No favourite

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

5 Did you write your passwords down at any time during this study?

- ○ Yes
- ○ No

6 Did you mentally rehearse your passwords between the two sessions in this study?

- ○ Yes
- ○ No

7 Other comments that may not have been addressed in the questions above?

Appendix E: Debriefing Form Part A – Interactive Web Site Study

(To be given at the end of Session 1)

This experiment is being conducted to examine the memorability and usability of passwords. It is hoped that this research will lead to an increased understanding of the use and security of passwords.
Please do not write your passwords down during this study. We are interested in testing your memory for the passwords when they are not recorded anywhere.
If you have any further questions regarding this research, please contact:

Nicholas Wright Principle Investigator Carleton University (613) 520-2600 Ext. 6317 nwright@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. Monique Sénéchal Chair, Carleton University Ethics Committee for Psychological Research Carleton University (613) 520 2600 ext 1155 monique_senechal@carleton.ca	Dr. Janet Mantler Chair, Department of Psychology Carleton University (613) 520 2600 ext 4173 psychchair@carleton.ca

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

Username: _____

Appointment: _____

Debriefing Form Part B – Interactive Web Site Study

This study is being conducted to examine the usability, practicality, and security of passwords, specifically in the context of recall and recognition conditions. Historical research suggests that a greater amount of information can be remembered when it can be recognized rather than recalled without any cues. However, passwords that capitalize on recognition are typically less secure. This situation poses a problem because as usability of a password increases, the level of security decreases. It is hoped that the assignment of passwords in both experimental conditions will allow us to compare and investigate the two mechanisms more deeply.
The time and effort you have spent as a participant in this study is very much appreciated!

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

Nicholas Wright Principle Investigator Carleton University (613) 520-2600 Ext. 6317 nwright@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. Monique Sénéchal Chair, Carleton University Ethics Committee for Psychological Research Carleton University (613) 520 2600 ext 1155 monique_senechal@carleton.ca	Dr. Janet Mantler Chair, Department of Psychology Carleton University (613) 520 2600 ext 4173 psychchair@carleton.ca

Appendix F: Interactive Web Site Study Reminder Notices to Participants
(To be sent by e-mail between visits)

Dear participant,
We would like to remind you to take a little time in the near future to login to the three sites mentioned during your first session using the three passwords that were assigned to you. In the event that you cannot login successfully, please take advantage of the password recovery tool available at each site.

Links to each of the sites:
Site 1
Site 2
Site 3

Thanks very much for your continued participation in the Interactive Web Site Study.
Sincerely,
Nicholas Wright

(To be sent by email before Session 2)
Dear participant,
We would like to remind you of your upcoming appointment to participate in an interactive web site experiment. You have a booked appointment for: _____
If you can no longer attend this appointment or would like to change times, please contact us at [illegible]
Thank you very much, and we appreciate your participation.

Nicholas Wright

Appendix G: Interactive Web Site Study Recruitment Poster

Interactive Web Site Study

2% or $20

Reward!

2 part study

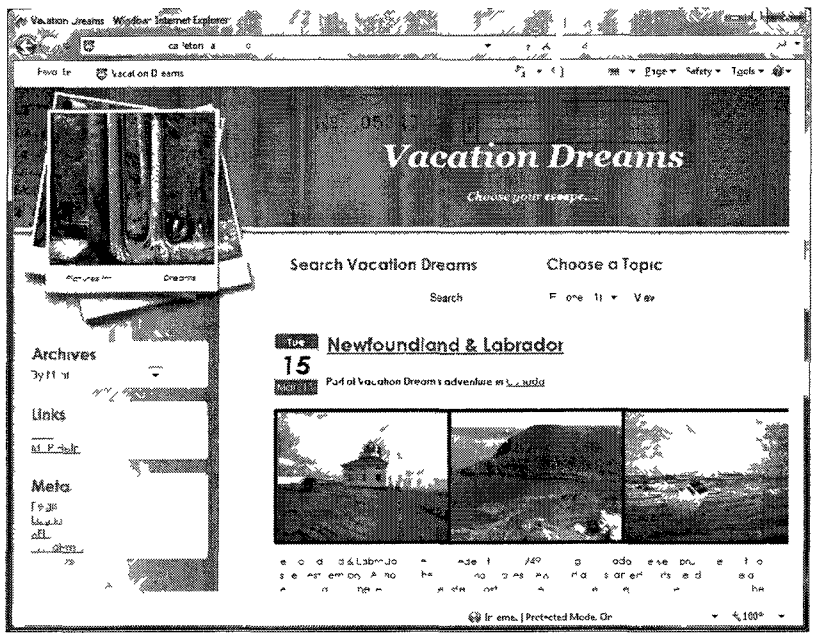
requires two visits, and brief online participation between appointments.

Ideal participants make regular use of the Internet and are familiar with the use of usernames and passwords to access web pages.

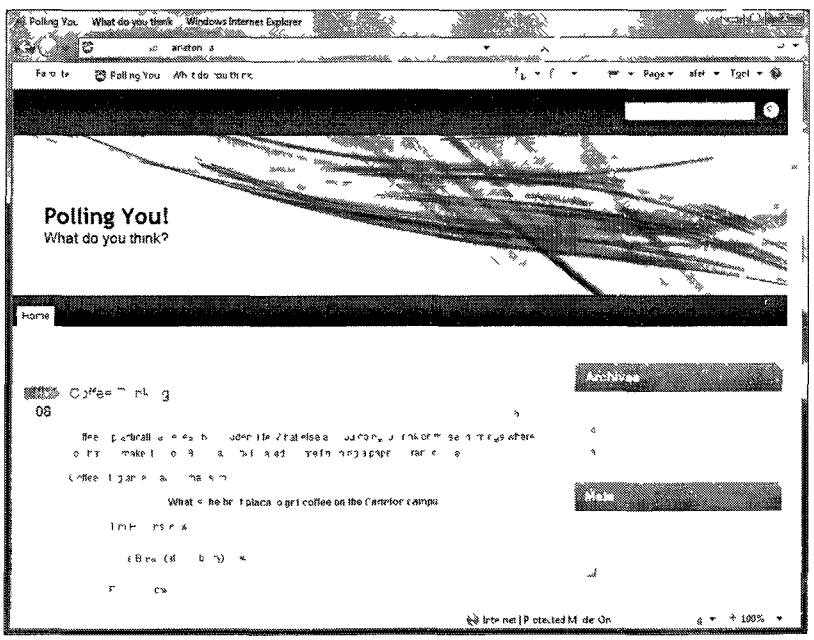
Interested parties can register via the SONA system, or contact:
Nicholas Wright, Principal Investigator at
nwright@connect.carleton.ca

Appendix H: Example Web Sites

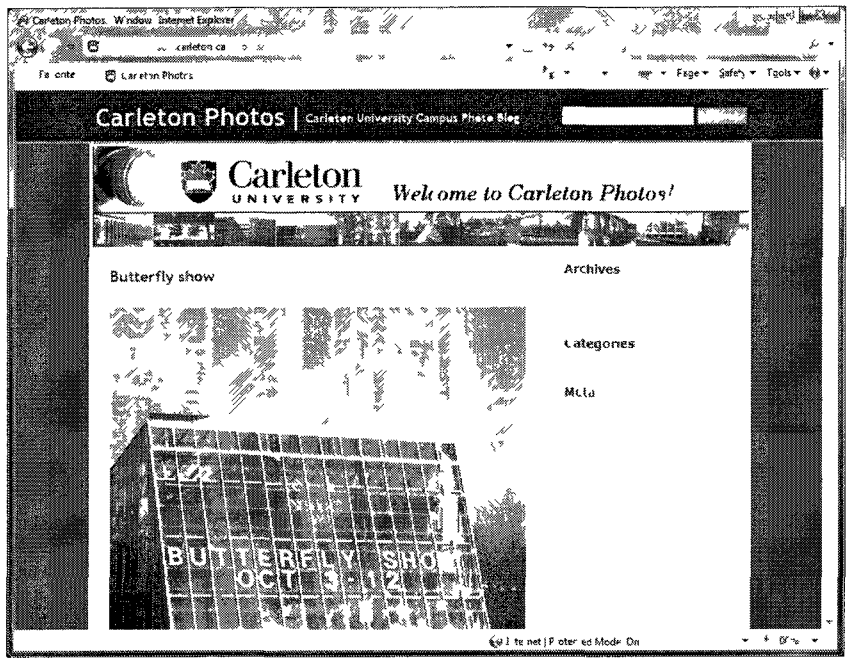
Vacation Dreams:



Polling You:



Carleton Photos:



Appendix I: Interactive Web Site Study Word List

across	almost	amount	animal	answer	attack
august	ballet	basket	beetle	before	belief
bitter	bottle	branch	breath	bridge	bright
broken	bucket	butter	button	camera	canvas
carpet	chance	change	cheese	chorus	church
circle	circus	coffee	cognac	collar	colony
colour	common	copper	cotton	credit	damage
danger	degree	design	detail	drawer	effect
eleven	empire	engine	enough	expert	family
father	feeble	female	finger	flight	flower
fourth	friday	friend	future	garden	growth
hammer	hollow	humour	insect	island	kettle
letter	liquid	little	living	locust	market
memory	middle	minute	monday	monkey	mother
motion	muscle	museum	narrow	nation	needle
nickel	normal	number	office	omelet	orange
parcel	patent	pencil	person	please	plough
pocket	poison	police	polish	porter	potash
potato	powder	prince	prison	profit	public
rabbit	reason	record	regret	reward	rhythm
rodent	school	second	secret	silver	simple
sister	smooth	sneeze	sponge	spring	square
sticky	stitch	street	strong	sudden	summer
sunday	system	theory	thirty	though	thread
throat	ticket	tongue	twelve	twenty	vessel
violin	weight	whisky	window	winter	yellow