MEMORABILITY OF ASSIGNED RANDOM GRAPHICAL PASSWORDS

by

ELIZABETH ANN STOBERT

A thesis submitted to
the Faculty of Graduate and Postdoctoral Affairs
in partial fulfillment of the requirements for the degree of

Master of Arts

in

Psychology

Carleton University
Ottawa, Ontario

Your file Votre référence
ISBN: 978-0-494-83110-6
Our file Notre référence
ISBN: 978-0-494-83110-6

Abstract

When allowed to select their own passwords, users often choose easily guessed passwords. Assigning random passwords removes this threat, but assigned passwords can be difficult to remember. Graphical passwords are an alternative form of authentication that use images for login and leverage the picture superiority effect for good usability and memorability. This thesis examines the memorability of random assigned graphical passwords, and compares them to text passwords. It also examines how different kinds of memory retrieval (recall, cued-recall, and recognition) affect the memorability of graphical passwords. A study of five password systems showed that participants were able to remember both graphical and text passwords for the duration of the study, but it was difficult to assess the memorability of the passwords because many participants wrote their passwords down. The usability of the schemes varied, but in general, it took longer for participants to login using password schemes that more leveraged recognition memory.

Acknowledgements

I would like to thank all of the following people for their assistance and support in the completion of this thesis. It would not otherwise have been completed.

My supervisor, Robert Biddle, for his enthusiasm, patience, and ongoing support.

The members of my committee: Andrew Patrick, Warren Thorngate and Anil Somayaji for their time and feedback.

Sonia Chiasson, for all of her friendship and mentorship.

My parents, for their encouragement, interest and patience.

Emily Miller-Cushon, for understanding that I had nothing else to talk about.

Contents

Introduction **1**
Usable Security . . . 3
Authentication . . . 4
Password Spaces . . . 5
Threat Models . . . 6

Graphical Passwords **7**
Picture Superiority Effect . . . 8
Memory Retrieval . . . 9
Types of Graphical Passwords . . . 12

Research Question **20**
Research Study . . . 20

Method **25**
Participants . . . 25
Apparatus . . . 25
Materials . . . 26
Procedure . . . 26

Hypotheses **28**

Analysis Plan **30**
Memorability . . . 30
Usability . . . 31
Exploratory Analysis . . . 31

Results **34**
Hypothesis 1: Memory Time . . . 34
Hypothesis 2: Resets . . . 37
Hypothesis 3: Login times . . . 40
Exploratory Analysis . . . 43
Security Analysis . . . 57
Questionnaire Data . . . 59

Discussion **63**
Results Summary . . . 63
Writing Passwords Down . . . 66
Picture Superiority Effect . . . 68
Memory Retrieval . . . 69
Passtiles Evaluation . . . 72

Conclusion **73**

References **75**

List of Tables

1 Theoretical password spaces for text passwords using varying character sets and password lengths. . . . 5
2 Password spaces for the configurations of the password systems used in the study. 24
3 Descriptive statistics for memory time (in hours). . . . 34
4 t-tests of memory time. . . . 36
5 ANOVA comparing memory time for the graphical password conditions. . . . 36
6 t-Tests of memory time. . . . 37
7 Descriptive statistics for password resets. . . . 37
8 Wilcoxon tests of password resets comparing the assigned text condition with each of the graphical password conditions. . . . 39
9 Kruskal-Wallis test of resets in the graphical password conditions. . . . 39
10 Wilcoxon tests of resets comparing the chosen text condition with each of the other conditions. . . . 40
11 Descriptive statistics for login times. . . . 40
12 Kruskal-Wallis test of login times. . . . 42
13 Pairwise Wilcoxon tests of login times using Bonferroni adjustment. . . . 42
14 ANOVA of memory time for all conditions. . . . 44
15 Descriptive statistics for failures. . . . 44
16 Kruskal-Wallis Test of failed password entries. . . . 46
17 Pairwise Wilcoxon tests of failed entry attempts using Bonferroni adjustment. . 46
18 Descriptive statistics for lasting successes. . . . 48
19 Kruskal-Wallis test of total number of lasting successes. . . . 48
20 Descriptive statistics for password interference. . . . 49
21 Kruskal-Wallis test of password interference. . . . 49
22 Descriptive statistics of ordered password entries. . . . 51
23 Pairwise t-tests of order repeats using Bonferroni adjustment. . . . 52
24 Linear regression of tile position on click number for Blank Passtiles. . . . 54
25 Linear regression of tile position on click number for Image Passtiles. . . . 54
26 Chi-squared test of password recordings. . . . 55
27 Frequency of character set use in Chosen Text passwords. . . . 58
J1 Descriptive statistics for memory time (in hours). (MTurk) . . . 102
J2 t-tests of memory time. (MTurk) . . . 102
J3 ANOVA comparing memory time for the graphical password conditions. (MTurk) 102
J4 t-Tests of memory time. (MTurk) . . . 104
J5 Descriptive statistics for password resets. (MTurk) . . . 104
J6 Wilcoxon tests of password resets comparing the assigned text condition with each of the graphical password conditions. (MTurk) . . . 104
J7 Kruskal-Wallis test of resets in the graphical password conditions. (MTurk) . . 105
J8 Wilcoxon tests of resets comparing the chosen text condition with each of the other conditions. (MTurk) . . . 105
J9 Descriptive statistics for login times. (MTurk) . . . 105
J10 Kruskal-Wallis test of login times. (MTurk) . . . 107
J11 Pairwise Wilcoxon tests of login times using Bonferroni adjustment. (MTurk) . 107

List of Figures

1 The Draw-a-Secret login interface, showing a password drawing. 13
2 The GrIDsure login interface, with the Personal Identification Pattern squares circled. 15
3 The PassPoints password creation screen, with the five click-points highlighted. 16
4 The Passfaces login screen. 19
5 Password creation interfaces for the three graphical password schemes used in the study. 21
6 A timeline of the three study sessions. 27
7 Distributions of memory time for each study condition. 35
8 Distributions of resets for each study condition. 38
9 Distributions of login time for each study condition. 41
10 Distributions of number of password entry failures for each condition. 45
11 Distributions of number of passwords remembered for the entire study for each study condition. 47
12 Distributions of instances of interference for each condition. 50
13 Spatial patterns and directionality of subsequent ordered clicks in the Blank Passtiles condition. 53
14 Spatial patterns and directionality of subsequent ordered clicks in the Image Passtiles condition. 53
15 Frequency of reported password recording. 55
16 Distributions of Likert scale responses to usability and security perception questions. 1 is strongly disagree and 10 is strongly agree. In 16(f), 16(g) and 16(i) the scales are reversed to show positive attributes consistently. 62
G1 The Vacation Dreams blog. 95
G2 The Polling You! blog. 95
G3 The University 101 blog. 96
J1 Distributions of memory time for each study condition. (MTurk) 101
J2 Distributions of resets for each study condition. (MTurk) 103
J3 Distributions of login time for each study condition. (MTurk) 106
J4 Frequency of reported password recording. (MTurk) 108

List of Appendices

Appendix A Ethics Application 80
Appendix B Informed Consent 84
Appendix C Demographics Questionnaire 87
Appendix D Interim Debriefing 89
Appendix E Post-test Questionnaire 91
Appendix F Debriefing 93
Appendix G Study Websites 95
Appendix H Recruitment Notices 97
Appendix I The Mechanical Turk Study 98
Appendix J Mechanical Turk Study Results 100

Introduction

One of the challenges of computer security is creating systems that are not only secure, but also usable. When the usability of a system is increased, it is possible for the security of the system to decrease, and vice versa. If a system is unusable, it will often be insecure since users will misuse or bypass the security mechanisms. Usable computer security (Cranor & Garfinkel, 2005) seeks to design "secure systems that people can use". One area of computer security is authentication, or the act of proving that someone is who they say they are.

Text passwords are a form of authentication that are widely used, but often insecure (Florencio & Herley, 2007). One of the major security vulnerabilities of text passwords stems from predictability in the way that users choose passwords. When users select passwords with predictable patterns such as dictionary words, leading capitalizations, or simple number substitutions, these patterns can be leveraged by attackers. By prioritizing often-chosen passwords, attackers can more efficiently guess passwords to break into users' accounts.

Assigning random passwords solves the security problems deriving from user chosen passwords. When passwords are randomly assigned, attackers cannot prioritize their password guessing attempts, and are left attempting to guess all possible passwords in no particular order. While text passwords can be easily assigned, they are difficult for users to remember, which leads to insecure coping behaviours, such as writing passwords down.

Graphical passwords are a proposed alternative to text passwords that have good security and usability properties (Biddle, Chiasson, & van Oorschot, in press). Instead of using text, graphical passwords ask users to complete some kind of image-based task to login. There are many different graphical password systems, but some popular systems ask users to draw a password image (Jermyn, Mayer, Monrose, Reiter, & Rubin, 1999), click different places on a picture (Wiedenbeck, Waters, Birget, Brodskiy, & Memon, 2005), or identify pictures of faces (Real User Corporation, 2004). Graphical passwords leverage the picture superiority effect (Paivio, Rogers, & Smythe, 1968), which says that humans remember images better than they remember textual information. In addition, different graphical password schemes leverage different methods of information retrieval. Recall-based graphical passwords ask users

to recreate a pre-set drawing to log in. Cued-recall passwords show users an image, and to log in, they must click the correct points on the image. Recognition-based graphical passwords present users with an array of images, and the user must choose the correct images to log in.

Graphical passwords have been shown to have good usability and memorability (Biddle et al., in press), but they can suffer from vulnerabilities of user-choice similar to those of text passwords. Studies of different graphical password systems (van Oorschot & Thorpe, 2008; Chiasson, Biddle, & van Oorschot, 2007; Chiasson, Forget, Biddle, & van Oorschot, 2009; Dunphy & Yan, 2007; Davis, Monrose, & Reiter, 2004) have shown that users tend to pick graphical passwords with exploitable patterns.

This thesis addresses the memorability of assigned graphical passwords. Are assigned graphical passwords more memorable than assigned text passwords? It also examines the effects of different kinds of memory retrieval on the memorability of graphical passwords.

The subsequent sections provide an introduction to usable security, authentication and password security, then outline current research on graphical passwords. The theoretical basis for graphical passwords is the picture superiority effect, which has been argued to suggest that graphical passwords will be more memorable than text passwords; this will be discussed first. Graphical password schemes leverage different kinds of memory retrieval, and different schemes claim different advantages relating to these differences, which may apply even to assigned random passwords. Specific examples of the different kinds of graphical passwords and their suitability to assigned random passwords is outlined. A detailed description of the research question and research study is given, along with an outline of the study methodology. The hypotheses and analysis plan are described, and the results of the study are given in detail. Finally, we conclude with discussion of the study results and outline some future work.

Usable Security

Usable computer security (Cranor & Garfinkel, 2005) attempts to increase the usability of computer security systems, while simultaneously maintaining or increasing the level of security the system supports. Users are often regarded as the weak link in security systems (Adams & Sasse, 1999), and many systems are designed without sufficient consideration of the user.

It is often difficult to create systems that are both usable and secure. For example, considering password authentication, it is possible to think of systems that are usable but insecure, such as a one-character password. Alternatively, long random-character passwords are said to be secure, but are slow for users to enter and difficult to remember.

The design of usable security software is complicated by the presence of *attackers*, or malicious users. Security tasks must be made easy for the user, but difficult for the attacker. Designing for usability is always challenging, but adding security to the task brings a number of distinct properties that must be acknowledged in the design of user interfaces for security (Whitten & Tygar, 1999):

1. Users are typically uninterested in security. Instead of being the user's main focus, security is a secondary task that they must complete in order to move on to a more relevant primary task, such as email or web browsing. They may be forced to enter a password to access a locked computer, but they are not interested in password entry in and of itself. Given that security is a secondary task, it is easy to understand why users avoid or bypass security mechanisms: they want to focus on their primary work.

2. Security systems are often complex and abstract because of the sophistication involved in techniques such as encryption and digital signatures. The security user interface must present information in a way that the user can understand and accurately manipulate. The complexity of security systems can also make it difficult to provide adequate and relevant feedback to the user about the state of the system. This further complicates the task of creating clear and understandable interfaces.

3. Since computer systems are constantly under the threat of attack, it is important that users not be able to make dangerous errors that can open the system to attack. Computer

security suffers from the "barn door" property. Similar to the proverb about not locking the barn after the horse has been stolen, it is important that computer systems *always* be secure. If a system is left unsecured for even a short period of time, there is no way to tell if it has been compromised because malicious software might remain hidden on the system. Thus, the system must always be secure, even as the user is learning and experimenting.

Authentication

One important area of computer security is *authentication*, or the act of proving that someone is who they claim to be (Menezes, van Oorschot, & Vanstone, 1996). Authentication differs from *authorization*, which refers to an individual having permission to complete some task, but the two often go hand-in-hand, with authentication being necessary for authorization.

There are three ways in which authentication may be granted (Menezes et al., 1996):

Something known: Information known by the individual is used to prove their identity. This category includes passwords, pass-phrases and PINs.

Something possessed: Also known as token-based authentication, this kind of authentication requires the possession of some kind of physical item. This category includes smartcards, as well as one-time password generators.

Something inherent: Also known as biometric authentication, this method of authentication depends on a biological or behavioural feature of the individual attempting to log in. This includes *physical biometrics*, such as fingerprints or retinal scans; and *behavioural biometrics*, such as typing patterns.

Authentication may involve multiple steps falling into different categories, such as at automated banking machines where the user is asked for both a bank card (something possessed) and something known (their PIN). The most common form of authentication is authentication by something known, which in everyday computing usually takes the form of text passwords.

Table 1: Theoretical password spaces for text passwords using varying character sets and password lengths.

Character set	Size	Length (n)	Size	Theoretical Password Space
All typeable characters	95	8	6.63×10^{15}	53 bits
All typeable characters	95	6	735,091,890,625	39 bits
Upper and lower cases, digits	62	8	2.18×10^{14}	48 bits
Upper and lower cases, digits	62	6	56,800,235,584	36 bits
Upper and lower cases	52	8	5.35×10^{13}	46 bits
Upper and lower cases	52	6	19,770,609,664	34 bits
Lower case letters, digits	36	8	2.82×10^{12}	41 bits
Lower case letters, digits	36	6	2,176,782,336	31 bits
Lower-case letters	26	8	208,827,064,576	38 bits
Lower-case letters	26	6	308,915,776	28 bits
Digits	10	4	10,000	13 bits

Password Spaces

Password security is usually measured by the size of the *theoretical password space*, or the total number of passwords that can possibly be created in the password system with a set of given parameters. For text passwords, the theoretical password space is usually calculated as 95^n, where 95 is the number of typeable characters on a US English keyboard, and n is the variable representing the password length. For a 6-character password, the theoretical password space would be $95^6 = 735,091,890,625$. For easier comparability, theoretical password spaces are usually expressed as exponents of base 2, and are traditionally referred to as being measured in "bits". Thus, $95^6 = 2^{39} = 39$ bits. Table 1 shows the theoretical password spaces for various character sets and lengths of text password.

While the theoretical password space indicates the number of possible passwords in a system, the *effective password space* measures the number of passwords typically selected by users. While the user can theoretically choose any of the 95 typeable keyboard characters in their passwords, users often limit their choices to smaller subsets of the keys, for instance, choosing only to include lower case alphabet letters. The effective password space is difficult to accurately estimate, since it varies with user preference and may not be evenly distributed across the theoretical password space (Weir, Aggarwal, Collins, & Stern, 2010).

Threat Models

There are two categories of attacks on password systems (Biddle et al., in press). *Guessing attacks* take place when an attacker attempts to break into an account by repeated guessing attempts of the passwords and *capture attacks* involve direct captures of the password by illicit means.

Guessing attacks include exhaustive searches and prioritized guessing attempts. A *brute force attack* occurs when an attacker systematically guesses every possible password. The theoretical password space gives an indication of the strength of the password system against a brute force attack. A *dictionary attack* takes place when an attacker uses a prioritized list of probable passwords (a dictionary) or prioritizing algorithm to attack an account. For text passwords, these are literally dictionaries, embellished with common misspellings and substitutions. Other exploitable patterns in text include leading capitalization and trailing punctuation.

When users are allowed to select their own passwords, they typically choose passwords which are easily guessed by dictionary attack (Florencio & Herley, 2007). Assigning random passwords removes this vulnerability, but assigned passwords can be difficult for users to remember, and cumbersome to enter, which limits their usability. This ensures that the effective password space can be equal to and distributed evenly across the theoretical password space. Currently, the majority of systems (websites, email, etc.) do not assign passwords, though some do.

Capture attacks include shoulder-surfing, phishing, and some kinds of malware, such as key-loggers. Shoulder-surfing takes place when an attacker attempts to learn a password by watching the user enter it and often involves the use of recording devices. Phishing is a social engineering attack where attackers attempt to trick users into sharing their login credentials on a fraudulent website. Malware is malicious software installed on the user's computer, and some varieties of malware capture all keyboard entries so that they can be searched for password information.

Graphical Passwords

Graphical passwords (Blonder, 1996) are an alternative to text passwords that use images instead of text for users to log in. Graphical passwords are intended to increase the usability of passwords by harnessing memory for pictures. Several surveys of graphical passwords are available, including ones from Suo, Zhu, and Owen (2005), Monrose and Reiter (2005), and Biddle et al. (in press).

The security of graphical passwords varies depending on the system, but graphical passwords have a few common security advantages and disadvantages (Biddle et al., in press). Graphical passwords tend to be resistant to recording. They are difficult to write down succinctly and accurately, and techniques such as screen captures can be awkward to store and access. Similarly, graphical passwords can be difficult to accurately and memorably describe to another person.

One threat to the security of most graphical password systems is shoulder surfing, or the ability of an attacker to learn the login information by watching the legitimate user. Shoulder surfing is a legitimate concern, but it is also a concern for text passwords, and tends to be a minimal risk for users entering passwords in physically secure usage environments. Social engineering attacks take place when malicious users attempt to trick users into sharing their login credentials, and the attacker uses these credentials to access the system. While there are variations between schemes, graphical passwords are generally more resistant to social engineering attacks.

One advantage of graphical passwords is their resistance to multiple password interference (Chiasson, Forget, Stobert, van Oorschot, & Biddle, 2009). Users often have passwords for many different accounts, which can lead them to resort to insecure coping mechanisms such as password reuse and writing passwords down. Some graphical password schemes are resistant to these behaviours, and other systems offer a visual cue or reminder to help users remember and distinguish their passwords.

Although it is clear from experimental work (Davis et al., 2004; Chiasson, Forget, Biddle, & van Oorschot, 2009; van Oorschot & Thorpe, 2008, in press) that users create graphical

passwords with patterns, it is less clear how a "dictionary" of these passwords should be developed. van Oorschot and Thorpe (in press) suggest the use of human computation to create attack dictionaries for graphical passwords.

Picture Superiority Effect

Graphical passwords are said to leverage the *picture superiority effect* (Paivio et al., 1968), or the finding that people have better memory for images than words. The picture superiority effect is cited as the reason that graphical passwords are more memorable than text passwords, but little work has investigated whether this effect of increased memorability extends to assigned random graphical passwords.

The picture superiority effect is seen in tests of both recall and recognition, but can be disrupted or eliminated. Nelson, Reed, and Walling (1976) showed that the effect is diminished when schematically similar pictures are shown, and Nelson, Reed, and McEvoy (1977) found that a rapid presentation rate eliminated the effect. Retrieval tasks may also be structured such that the picture superiority effect is minimized or reversed (Weldon & Roediger, 1987; Weldon, Roediger, & Challis, 1989).

Several explanations for the picture superiority effect have been proposed. Paivio's *dual coding theory* (1971) postulates that the brain has separate mechanisms for remembering imaginal information (such as objects, images and events) and for remembering verbal information (both spoken and written). The picture superiority effect is speculated to be due to the dual coding that occurs when people remember images. Not only are the images encoded visually and remembered as images, they are also translated into a verbal form (as in a description) and remembered semantically.

Other explanations for the picture superiority effect speculate that images have implicit properties that make them more memorable. These explanations were later collectively identified by Mintzer and Snodgrass (1999) as the *distinctiveness account.* Nelson et al. (1977) proposed the sensory-semantic model, and argued that the picture superiority effect occurs because, although words and images share identical semantic codes, images are accompanied by more distinct sensory codes, allowing them to be more easily accessed. This theory is sup-

ported by evidence that visual similarity in images leads to decreases in the picture superiority effect (Nelson et al., 1976).

The levels-of-processing approach (Craik & Lockhart, 1972) breaks with traditional multistore models of memory and proposes that the endurance of information in memory has to do with the quantity and quality of processing and encoding it undergoes in memory. Applying this framework, Nelson and Reed (1976) found evidence that the picture superiority effect is related to the different levels of processing applied to images and words.

Mintzer and Snodgrass (1999) evaluated differences in recognition memory for studied words and pictures when tested in their studied or unstudied forms, and found support for the distinctiveness account. They measured the *form change cost*, or the difference in recognition performance between items tested in their studied form and items tested in a different form. They found a larger form change cost for pictures than for words, contradicting the predictions of dual-coding theory and supporting the hypothesis that the sensory and semantic features of images are uniquely encoded in memory.

Memory Retrieval

Different graphical password schemes leverage different types of memory retrieval. The differing kinds of retrieval may affect not only memory, but other factors, such as the time to login, or the ease of use.

Recall and recognition are processes of retrieving information from memory. Framed in early work as opposite memory tasks, *recall* is the process of remembering a specific focus when the context is provided, whereas *recognition* is the process of remembering the contextual information when the focus is provided (Hollingworth, 1913). Recall can be divided into *cued* recall, where a cue provides assistance in retrieval of the correct memory, and *free* recall, where no support is given.

Much research has focused on the relationship between recall and recognition memory, and several theories have been proposed to explain how they work.

Strength theory (Wickelgren & Norman, 1966) speculates that recall and recognition involve the same memory task, but that recognition requires a lower threshold of strength, and

is thus easier. Dual-threshold theory (Watkins & Gardiner, 1979) speculated that the trace strength of a memory must be above a certain threshold for the memory to be either recalled or recognized, and that the recognition threshold was lower than the recall threshold. Strength and threshold theories draw from evidence that variables such as timing (of presentation, of retention, etc.) affect recall and recognition memory similarly. However, studies have shown that variables such as intentionality of learning affect recall differently than they affect recognition, which does not support strength and threshold theories.

One of the most prominent theories of retrieval is *generate-recognize theory* (Anderson & Bower, 1972). This theory posits that retrieval is a two-step process, consisting of both generation and recognition phases. In the generate phase, long term memory is searched, and a list of candidate words is formed. Then in the recognize phase, the list words are evaluated to see if they can be recognized as the sought out memory. The model assumes that words occupy fixed positions in memory, with one (or occasionally, a small number of) meaning(s). When a word is encountered, a "tag" is appended to the word memory, giving some description of the situation of the encounter. In the recognition phase, these tags are assessed to determine if the item is correct.

Generate-recognize theory explains some of the differences between recognition and recall memory (Watkins & Gardiner, 1979). Since recognition memory does not utilize the generation phase, it is faster and easier to perform. The theory also explains the benefits of cueing on memory retrieval. A cue can help not only in generating a relevant candidate list, but also in recognizing the appropriate word from that list.

Although the generate-recognize model explains a number of experimental findings, there are also findings that contradict the model, or whose results are not accommodated by the model (Watkins & Gardiner, 1979). Most notably, the theory has difficulty explaining the success and failure of some kinds of cueing. Studies have shown that when the cue for a studied list item is changed (eg., "sail" vs. "gravy" for boat), subjects have a harder time recalling the appropriate word. Conversely, in studies of recognition, non-studied cues can cue the retrieval of unrecognized words.

In reaction to research about the effects of unlearned cues on recall memory, Tulving

and Thompson (1973) posited *encoding specificity theory.* This theory states that only stored information can be retrieved, meaning that only the information processed at the time of storage can later be used as retrieval cues. If semantic information about a word is processed at the time of learning, then that information can successfully be used to cue memory. Thus, the word "table" can only be used to cue memory of the word "chair" if the subject encodes the semantic information linking the two objects at the time of encoding. According to encoding specificity theory, if the word "violet" is encoded in the context of a flower name, it will not be successfully cued with the suggestion of a colour name.

Tulving and Thompson found support for their model in a study where participants were asked to remember a set of target words, where each word was accompanied by a specific input cue so that participants would encode the target words in relation to the cues. Then, participants were given a set of cues which were not in the original list and asked to generate a set of words through free association, before being asked to identify which of their generated words could be found on the original list. Finally, participants were given a list of the original input cues, and asked to recall the original list words. Participants were largely unable to recognize their original words on the list generated by themselves, but were able to recall original words when presented with the original cues, indicating that the encoding information affects memory.

One of the major differences between generate-recognize theory and encoding specificity theory is the assumed level of complexity of the retrieval process (Ellis & Hunt, 1989). Generate-recognize theory is a two-process theory, indicating that retrieval is a different process in recall and recognition. Since recall requires both the generation and decision phases, it is fundamentally more complex than recognition, which requires only the decision phase. In contrast, encoding specificity theory assumes that retrieval is an automatic and uncomplicated process, and the complexity occurs in the encoding task. Evidence exists to support both theories, and neither has been conclusively supported.

Recognition is almost always found to be superior to recall (Haist, Shinamura, & Squire, 1992), but a few studies (Tulving, 1968; Watkins, 1974) have shown recall to be sometimes superior to recognition. These studies have involved changing the context in which participants

were asked to remember paired words. In the learning condition, participants memorized word pairs that formed compound words, such as "air / port" or "home / stead". In the recall conditions, participants were better at recalling the second word (when presented with the first) than they were at recognizing the words from a list.

Cued-recall occurs when retrieval is aided by the presence of a cue. Different cues can be more or less effective, and it is not always clear what will make a good cue to memory. *Associative-strength theory* (Ellis & Hunt, 1989) says that a cue is effective if it has previously occurred with the remembered event in the past. The more frequently the events have occurred together, the higher the associative strength and the more effective the cue. Associative-strength theory assumes that memory is structured as a network that connects all items in memory. Items in memory with stronger ties between them make better cues, and the strength of the tie is increased by the frequency with which the two items occur together. In contrast, encoding specificity theory (Tulving & Thompson, 1973) says that the most effective cues are the cues that are present at the time of remembering.

Types of Graphical Passwords

Graphical password schemes can be divided into three categories (De Angeli, Coventry, Johnson, & Renaud, 2005), based on the kind of memory leveraged by the scheme:

Drawmetric: In these systems, the user is asked to reproduce a drawing on a grid. This category of graphical passwords is also known as recall-based. Example schemes include Draw-a-Secret (Jermyn et al., 1999) and Pass-Go (Tao & Adams, 2008).

Locimetric: In these passwords, the user is asked to accurately click on points on an image. These schemes are also referred to as cued-recall or click-based graphical passwords. Example schemes include PassPoints (Wiedenbeck et al., 2005), and Persuasive Cued Click-Points (Chiasson, Forget, Biddle, & van Oorschot, 2008).

Cognometric: In these schemes, the user is asked to recognize and identify images belonging to their set of password images from a set of distractor images. This category of graphical passwords are also known as recognition-based. An example password scheme

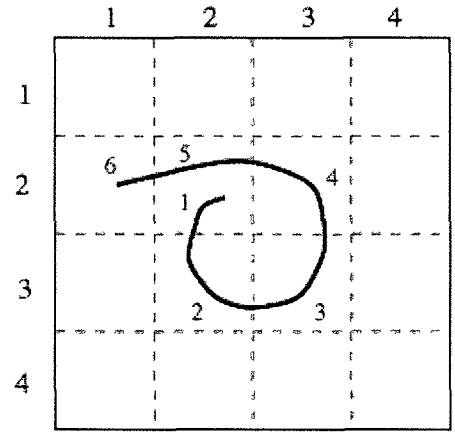


Figure 1. The Draw-a-Secret login interface, showing a password drawing.

is PassFaces (Real User Corporation, 2004).

The categorization of graphical passwords is based on the kind of memory retrieval harnessed by each password type. Cognometric graphical passwords employ recognition memory, and involve *recognizing* the correct password from a set of distractors. Drawmetric passwords must be *recalled* by the user, without any cues or reminders. Locimetric passwords also rely on recall memory, but the background image provides users with a cue to the password memory.

A variety of different graphical password systems exist, with different advantages and capabilities. A few relevant graphical password systems in each of the categories are outlined below.

Drawmetric. Draw-a-Secret (DAS) is a drawmetric graphical password scheme proposed originally by Jermyn et al. (1999). In DAS, the user is presented with a blank grid, and is asked to draw a secret picture on the grid (Figure 1). To log in, the user is asked to reproduce the same drawing on the grid. To log in successfully, the drawing must go through the correct grid squares in the correct order, and strokes must begin and end in the correct grid squares. Pass-Go (Tao & Adams, 2008) is a drawmetric graphical password system similar to DAS. In Pass-Go, the user is presented with a grid and draws a figure on that grid, but the lines are "snapped" to the grid coordinates.

The original work on DAS (Jermyn et al., 1999) cited its theoretical password space as

58 bits (assuming passwords with stroke count 12). van Oorschot and Thorpe (2008) noted the human propensity for mirror-symmetry, and investigated how the selection of symmetrical passwords would affect the password space of DAS. They suggested and followed a predictive method for modelling user-choice in graphical passwords. Their process involved identifying the user's tasks and the accompanying memory load, then determining relevant information about the user's ability to handle the demands on memory. Based on this information, they identified password complexity properties and used those properties to model classes of memorable passwords. Finally, they estimated the size of those classes to produce estimates of *weak password subspaces*, or generable subsets of the password space that contain easily guessed passwords.

As examples of their methodology, van Oorschot and Thorpe (2008) examined the sets of mirror-symmetrical passwords and passwords with small numbers of stroke counts in DAS. They found that these properties created weak password subspaces with sizes between 31 and 41 bits. Although they did not conduct a user study of DAS to confirm that users would choose passwords in these subsets, other studies of DAS (Chiasson et al., 2010) confirm the tendency for users to select passwords with short stroke counts.

Although most recall-based graphical password systems have been based in a grid-drawing exercise, there is no particular reason that a drawing task is necessary. In any graphical password system where no cue is given, the password system will leverage free recall. GrIDsure (GrIDsure, 2011) is a commercially deployed drawmetric graphical password system where the user selects a Personal Identification Pattern (PIP), or an ordered sequence of four grid squares on a password grid. At login, each grid square is filled in with a digit, and the user must enter the sequence of digits corresponding the grid squares in their PIP. The digit distribution in the squares changes at every login, so that the entered PIP is different every time. Figure 2 shows the GrIDsure login interface. These one-time passwords protect the scheme against simple capture attacks. Weber (2006) states the theoretical security of GrIDsure as 19 bits, but Bond (2008) notes that the scheme is susceptible to several vulnerabilities, including issues stemming from poorly chosen passwords, and that as such, the effective password space is probably small.

The complexity and granularity of existing drawmetric password systems makes assigning

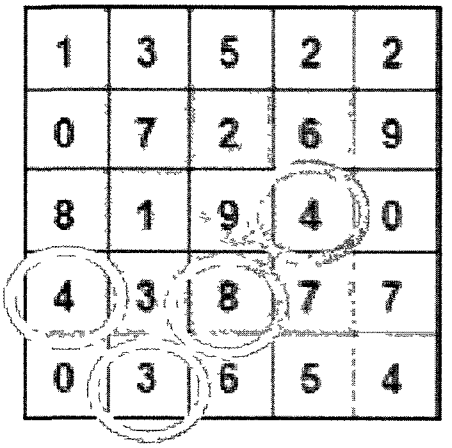


Figure 2. The GrIDsure login interface, with the Personal Identification Pattern squares circled.

passwords difficult. Clearly conveying the appropriate details is problematic in a system like DAS, where there are many subtle nuances to password entry and they are difficult to convey to users.

Locimetric. Locimetric, or cued-recall, graphical passwords present the user with a visual cue to help them recall their secret password. Different cued-recall graphical systems utilize different methods of password entry but all locimetric graphical password systems provide users with a visual cue to their password. These cues help to more easily recall and distinguish their passwords.

PassPoints is a click-based locimetric graphical password scheme proposed by Wiedenbeck et al. (2005). In PassPoints, a user is presented with a single image, and is asked to choose a number, n, of ordered click-points on that image. To log in, the user is required to click on the same points (within a given tolerance region, r) in the original order. Cued Click-Points (Chiasson, van Oorschot, & Biddle, 2007) is an extension of PassPoints, in which users are shown five images, and asked to choose one click-point on each image. To log in, they must click on the same points on the same images (within a predefined tolerance region). Because CCP gives individual cues for each click-point, it has usability advantages over PassPoints. Studies of PassPoints (Chiasson, Biddle, & van Oorschot, 2007) and CCP (Chiasson, van Oorschot, & Biddle, 2007) showed that different users tended to choose click-

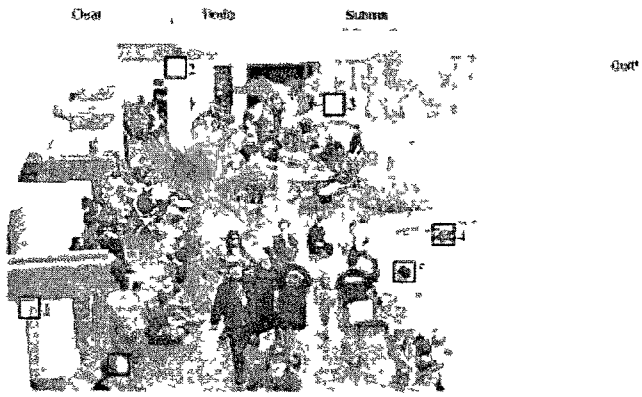


Figure 3 The PassPoints password creation screen, with the five click-points highlighted.

points in similar areas of the images, known as "hot-spots". Persuasive Cued Click-Points (PCCP) (Chiasson et al., 2008) added a persuasive viewport to the password creation phase. The viewport helped users distribute their click-points more evenly across the images, and minimized the hot-spot problem seen in PassPoints and CCP.

Two studies of PassPoints by Chiasson, Biddle, and van Oorschot (2007) showed PassPoints to be usable, with reasonable login times and success rates. The studies also showed that users were very accurate in re-targeting their click-points during login. Using data from the same studies, van Oorschot and Thorpe (in press) examined the prevalence of hot-spots, and found that the passwords in created in the study were extremely susceptible to dictionary attack. Chiasson, Forget, Biddle, and van Oorschot (2009) analyzed data from the lab and field studies of PassPoints (as well as data from further refinements of the PassPoints system) for password patterns including click-point distribution, segment lengths (i.e. click-point proximity), angles and slopes, and shapes. They found evidence of more click-point patterning than could be expected by chance, and that PassPoints passwords tended to contain both hot-spots and patterns that could potentially be leveraged by attackers.

Extending previous work on Draw-a-Secret, Dunphy and Yan (2007) investigated the effects of adding background images to the grids used in DAS passwords. They created a system called Background Draw-a-Secret (BDAS) where the user was asked to choose a background image for their password before being asked to draw a password on the grid. Dunphy and

Yan hoped to introduce the advantages of cued-recall to the DAS password scheme. They also identified other advantages that an image background could give users of DAS. These included speculation that the presence of a background image would help users more accurately situate their drawings on the grid. They also noted that the presence of a background image might not lead to the hot-spot effects seen in work on click-based locimetric passwords, since users of DAS have the freedom to highlight an area of interest in multiple ways, including drawing around it, through it, or beside it.

To investigate the effects of a background image on DAS passwords, Dunphy and Yan performed two paper-based user studies, where users created and re-entered DAS passwords both with and without background images. The study investigated password complexity by examining the number of strokes in the created passwords as well as the password length. They found that the password length for participants in the BDAS condition was significantly longer than those in the DAS group. The study investigated password complexity by examining the number of strokes in the created passwords as well as the password length. They found that the password length for participants in the BDAS condition was significantly longer than those in the DAS group.

One disadvantage of existing locimetric graphical password systems is the difficulty in assigning passwords. Multiple studies (Chiasson, Biddle, & van Oorschot, 2007; Chiasson, Forget, Biddle, & van Oorschot, 2009; van Oorschot & Thorpe, in press) have shown that allowing users to choose their own passwords leads to security vulnerabilities. Assigned passwords would remove these vulnerabilities, but the precise input required in PassPoints, CCP, PCCP and BDAS creates difficulties in communicating assigned passwords to users. However, there is no reason that a visual cue could not be used in a password system that required less precise input.

Cognometric. Cognometric graphical passwords work by presenting a grid of images, where one image belongs to a known set of "password" images, and the other images are distractors, and the user must correctly choose the password image to authenticate. Cognometric graphical passwords leverage recognition memory by explicitly displaying all possible choices

to the user, and expecting them to recognize the correct option.

In the commercial graphical password system PassFaces (Real User Corporation, 2004), the user has a portfolio of face images that make up their "password" and to login, they are asked to select one of their portfolio faces from a grid of distractor faces (Figure 4). The number of distractor faces is typically set to 8, and the number of images that must be identified to log in is usually set to be 4. Only one known face image is used in each grid, and to prevent information leakage to attackers, the distractor faces are always the same, though their placement is shuffled. Passfaces attempts to leverage the human ability to remember and quickly recognize faces. The theoretical password space of Passfaces is calculated as n^k where n is the number of faces in the displayed grid, and k is the number of presented grids. In its standard configuration of n and k ($n = 9$ and $k = 4$) the password space is approximately 13 bits, which gives security comparable to a 4-digit PIN.

Any kind of image may be used in a recognition-based password scheme. Hlywa (2011) compared house images and object images to face images, and found that face images were no more memorable for users than other kinds of images. As well, the study showed that the security of these schemes may be viable at higher security configurations created by adjusting the number of images in the password portfolio, and the number of distractor images in each presented grid.

Psychological studies have shown that people tend to form similar assessments of beauty, even across cultures. Studies have also shown that people have an easier time recognizing faces from their own group. Davis et al. (2004) conducted a study examining user choice in Passfaces passwords, and found that these biases affected the faces chosen for use in passwords. They ran a user study in which participants were asked to choose their own set of faces. They found three effects: that of gender, that of attractiveness and that of group. They organized the face images in their study into several categories. For gender, male or female; for beauty, attractive or typical; and for group, "black", "white" or "Asian". They examined the effects of these variables on the security of the passwords chosen. In the study, three university classes created and used Passfaces passwords over the course of a semester to access their class materials online.

Figure 4 The Passfaces login screen

The results of the study showed that gender, attractiveness and group all played biasing roles in password selection. They found that both men and women selected female faces significantly more often than male faces, and that men almost always chose attractive female faces. Their analysis showed that if a user is known to be male, an online dictionary attack will succeed in two guesses 10% of the time. If the user was Asian and their gender was known, then 10% of passwords could be guessed in six attempts. Previously, users of Passfaces were allowed to choose their own passwords, but as a result of this research, Passfaces began using assigned passwords.

Recognition-based passwords are easily assigned by showing the user their set of password images at password creation. They can be time-consuming for users to log in, since users have to search the shuffled array of distractor images for their password image. This process becomes slower as the security is increased because of the increased number of images.

Research Question

Assigned passwords have long been thought to be unmemorable and unusable. However, security issues such as guessing attacks that arise as a result of user-chosen passwords are non-trivial and exploitable by attackers. Although graphical passwords have been shown to be usable, they too remain vulnerable to the issues of user-choice. One method of avoiding these issues is to assign random passwords.

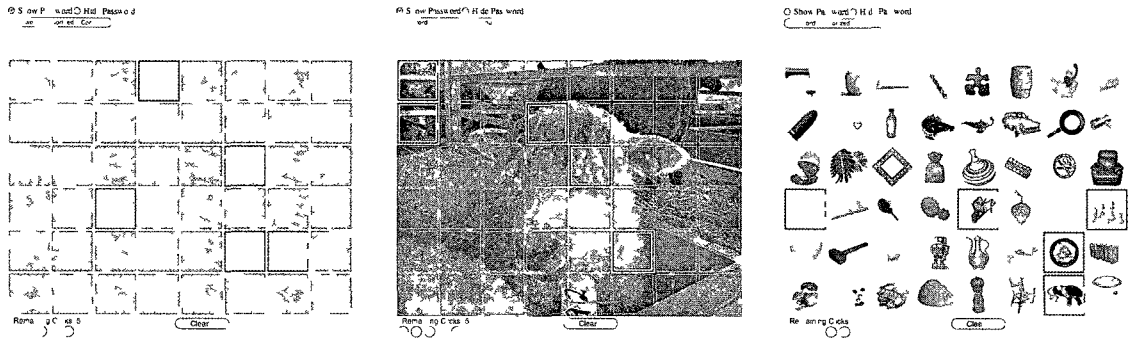
Though they are often dismissed as unusable, little work has investigated the usability of assigned passwords, or addressed how their memorability can be improved. This study will investigate whether assigned random passwords can be made memorable and usable through graphical passwords. This thesis concerns how different methods of information retrieval (recall, recognition and cued-recall) affect the memorability of assigned graphical passwords. Which of the retrieval methods is best for assigned random passwords? Are assigned graphical passwords more memorable and usable than assigned text passwords?

Research Study

To investigate this question, a between-subjects study was conducted. Participants were randomly assigned to one of five conditions, where the password scheme varied by condition. The study took place in three parts. Participants began the study by learning to use the password system and creating their passwords in the lab. In the second part of the study, they were asked to enter their passwords several times in an online setting. In the third part of the study, participants returned to the lab and were once again asked to enter their passwords. Participants had passwords for three different websites, and the study took place over one week.

The five study conditions were: Blank Passtiles (BPT), Image Passtiles (IPT), Object Passtiles (OPT), Assigned Text (AST), and Chosen Text (CHT). Three conditions used a general kind of graphical password scheme: *Passtiles.*

Passtiles (Figure 5) is a new graphical password system created for this study. It combines features of DAS, PassPoints and Passfaces to be able to easily assign passwords to users. Passtiles presents the user with a grid of password tiles and the user's password consists of five



(a) Blank Passtiles (b) Image Passtiles (c) Object Passtiles

Figure 5 Password creation interfaces for the three graphical password schemes used in the study

password tiles, which can be either assigned or chosen by the user. To log in, the user must click on the correct password tiles in any order.

Blank Passtiles (Figure 5(a)) is a variant of Passtiles that uses a grid with a blank background. Having a blank background makes password retrieval a pure recall task for the user, similar to DAS.

Image Passtiles (Figure 5(b)) is a variant of Passtiles that superimposes the grid of password tiles over an image. Having the grid of password tiles over an image lets users remember their password not only in relation to the background image but in relation to the grid of password tiles. Similarly to PassPoints, Image Passtiles takes advantage of cued-recall memory, and the image shown provides users with a cue to help them remember their password tiles.

Object Passtiles (Figure 5(c)) works similarly, but in each password tile, an object image is shown, creating a grid of smaller object images. The password consists of a set of objects that the user must click on to login (in no particular order). The same set of object images is always shown, but they are shuffled at every login. As in Passfaces, Object Passtiles takes advantage of recognition memory, where the user's password objects are always shown on the screen, but they must rely on recognition memory to find them in the shuffled grid.

The goal for Passtiles was to create a password system that could be used as a common framework to compare recall, cued-recall, and recognition graphical passwords. We needed a system that permitted random assignment of passwords, and had high learnability. Learnabil-

ity was emphasized because we wished the schemes to be viable in a real-world situation, and not require participants to undergo extensive training. Although existing graphical password systems leverage different kinds of memory, the schemes vary in appearance and functionality, and it would have been difficult to compare the effects of memory retrieval without other confounds. In addition, the complexity and fine detail of many existing schemes make it difficult to assign passwords and communicate them clearly to users. The complexity of existing systems also presented confounds in the form of learnability of different schemes. Passtiles needed to be comparable, flexible, easily learnable, and present assigned passwords clearly.

We chose to design the password system around a grid and tile model. The basic concept for the system was that it would present a grid of password tiles, and the password would consist of a subset of those tiles. The design of Passtiles allows for flexibility in display and password space. Passtiles allows the size of the presented display to be easily modified, and the dimensions of the grid to be modified separately. The length of the passwords can be set to any number of tiles.

An additional design element of Passtiles was the practice mechanism. Since we wished users to be able to actively participate in the memorization process, we designed a practice mechanism where users could practice entering their passwords during the create phase. In this phase, users could click on their password tiles, and would receive feedback about the correctness of their attempt. We added buttons to hide and show the assigned passwords, allowing users to practice entering their passwords without seeing them displayed, while still allowing users to see the password at any moment.

Passtiles was written in JavaScript, using the Raphaël library (Baranovskiy, 2010). JavaScript was chosen for its accessibility in all major browsers and operating systems, and the Raphaël library allowed easy implementation of the visual and image elements of the system. The system used a model-view-controller architecture (Burbeck, 1987) that allows the same underlying system to be used for all the variants of the scheme. It uses HTML elements to present parts of the user interface. Passtiles is deployed as part of the existing MVP framework (Chiasson et al., 2010) and uses server-side PHP scripting to communicate with MVP. MVP is an implemented framework for running user studies that allows different password systems

to be used on the same websites. MVP logs data about password system use, allowing for detailed analysis.

The remaining two study conditions used text passwords and provided comparison to a traditional password form. The Assigned Text condition gave a comparison to randomly assigned passwords, and the Chosen Text condition provided a controlled comparison to examine the level of security that users choose for themselves. The text password systems used in the study followed standard practices, but the assigned text password system included a practice mechanism with hide and show functionality to make it comparable to the assigned Passtiles passwords. For all of the password systems used in the study, account creation included two steps: password creation, where the user was assigned or selected their account password; and password confirmation, where the user confirmed their account password. Whether the password was assigned to or chosen by the user, this process is referred to as "password creation".

In order to a create a valid comparison, the parameters of the password schemes were set so that all five conditions had approximately equal theoretical security. Florencio and Herley (2010) suggest that 20 bits of security is sufficient for everyday computing, and we chose to use this as a guideline for the security settings in the study. For the Passtiles passwords, a grid of 8×6 tiles, and passwords of length 5 give a password space of 21 bits. Text passwords of length 4 using only lowercase letters and digits have a theoretical password space of 21 bits. Since the passwords were randomly assigned in the graphical password and Assigned Text conditions, the effective password space was equal to the theoretical password space. In the Chosen Text condition, the effective password space was likely considerably smaller than the theoretical space. Table 2 shows the password spaces in bits for the systems used in the study. In a pilot study of assigned text passwords where participants were asked to remember a total of 17 characters in 4 passwords, we found that participants were not able to remember passwords of this length, and resorted extensively to repeated resetting of passwords. Thus, we hoped that a slightly smaller load would be more manageable for users (12 characters in 3 passwords).

An ongoing concern in password studies is *ecological validity*, or the realism of the study situation. When studying passwords, it can be difficult to know whether people exhibit the

Table 2: Password spaces for the configurations of the password systems used in the study.

Password system	Configuration	Size	Theoretical Password Space
Blank Passtiles	8×6 grid, length 5	1,712,304	21 bits
Image Passtiles	8×6 grid, length 5	1,712,304	21 bits
Object Passtiles	8×6 grid, length 5	1,712,304	21 bits
Assigned Text	36 characters, length 4	1,679,616	21 bits
Chosen Text	36 characters, length 4	1,679,616	21 bits

same behaviour in the study that they would in real life. In an effort to elicit realistic behaviour, we chose to study passwords in the context of website use. The MVP framework (Chiasson et al., 2010) was used to implement the password systems on real websites. MVP allows different password systems to be implemented on the same websites and thus compared under identical conditions. The websites used in the study were configured to have dramatically different appearances, and used the MVP framework to allow the use of graphical password systems. The websites used in the study were created, hosted, and maintained by the experimenters.

The independent variable in the study was the password system used. The effect should indicate the usability of the scheme including both memorability and other relevant factors. The dependent variables were therefore the average length of time that a password was remembered, the number of password resets, and the time to login. To measure password memorability, we measured the *memory time*, or the average length of time that a participant remembered their password. For each account, it was measured as the greatest length of time between a password creation and the last successful password login (using the same password). The MVP system allows users to reset their passwords without experimenter intervention when forgotten. In a pilot study, we found that participants were repeatedly resetting their passwords rather than attempting to remember them. To subtly discourage this behaviour without introducing a monetary penalty, we introduced a 5 minute delay to the reset system and warned participants that it would take a few minutes for their passwords to be reset. The number of password resets per account was a dependent variable. Login time was measured from the time that the password entry window appeared on screen until the website verified the entry attempt as successful.

Method

Participants

Participants were recruited from both the university community and the wider community. To recruit from the university community, we put up posters around campus, and recruited via the psychology department online study system (SONA). For the wider community, we recruited participants using off-campus posters, mailing lists, and word of mouth. Our participants had to have regular access to a computer with internet, and be accustomed to entering usernames and passwords to access websites. They were also required to be aged 18 or over.

There were 81 participants in the study, with 16 participants in every condition except Chosen Text, which had 17 participants. Participants ranged in age from 18 to 62, with a median age of 24. 45 participants were female, and 35 were male (one person did not answer the question). 53 participants were students from a broad range of degree programs and levels. The remaining participants were occupied in a variety of fields including teachers, administrative assistants, writers, engineers, accountants, and analysts. None of the students were studying computer security, and no one worked in the field of computer security.

Fifteen participants reported having previously used a graphical password system, and all were referring to systems where the website provided an image for labelling as part of the authentication process (such as on the ING Direct login, where a user-specific image is shown in an attempt to assure the authenticity of the site). 8 participants reported having previously participated in a website study, and 5 participants reported previous participation in a password study.

Apparatus

The study took place in three parts, the first and third taking place in the lab, and the second taking place at home. For the in-lab sessions, the apparatus used in the study included a desktop computer running Windows 7, with a 17-inch monitor and US English keyboard. The websites used in the study were hosted online, and were accessed in the lab session using

the Mozilla Firefox web browser (version 3.6). All questionnaires administered in the study were administered online, using LimeSurvey (Schmitz, 2011) software. In the at-home session, participants used their regular computers to access the study websites and complete the study tasks.

Materials

The materials used in the study included consent (Appendix B) and debriefing (Appendices D and F) forms, as well as two questionnaires: a demographics questionnaire (Appendix C), and a post-test questionnaire (Appendix E). The consent form gave a brief explanation of the study, the timeline, and payment details. It also provided contact information, information about data anonymization and an explanation of the withdrawal policy. The interim debriefing form provided a few explanatory details about the study, as well as contact information. It informed participants that more information would be provided at the completion of the study. The final debriefing form gave a detailed explanation of the study, and- provided contact information should the participant have had further questions about the study.

The three websites used in the study were created, hosted and maintained by the experimenters. The content on each website focused on a different topic and contained non-offensive and non-controversial material. The three websites each had a distinct look and feel. Appendix G shows screenshots of each website.

The Object Passtiles graphical password scheme used a set of 375 images obtained from the stock.xchng (2011) photo website. The Image Passtiles password system used a set of 350 images, obtained with permission from personal collections, and from free photo websites. The images used were chosen for their clear focus, and high level of visual interest distributed across the image. Figure 5 shows screenshots of each password system.

Procedure

The study took place in three sessions, over the period of one week. Figure 6 shows a timeline for the three study sessions.

Session One	Session Two						Session Three
Day 0	1	2	3	4	5	6	7
Lab session	email		email			email	Lab session

Figure 6. A timeline of the three study sessions.

Session One. The first session took place in the lab and lasted approximately one hour. Before beginning, participants were given a brief explanation of the study, and were asked to read and sign the informed consent (Appendix B). The participant completed a brief training module on the password system, in which they learned how the password system worked and were allowed to practice using the system. They were introduced to each of the three websites in the study, and for each one created and confirmed a password before logging in. Next, they took a brief break from the websites to complete the demographics questionnaire (Appendix C). The demographics questionnaire asked participants about their age, gender, occupation, level of familiarity with internet use, and previous participation in similar studies. The participant then returned to the websites and was asked to complete a short task (such as commenting on an article) on each site. These tasks necessitated logging in to each website. Finally, the participant received an interim debriefing form (Appendix D), containing a few details about the study, a reminder about the upcoming sessions, and contact information.

Session Two. In the second session, the participant was sent three notification emails. Each email asked them to complete one task on each website. The emails were sent on the first day after session one, the third day after session one, and the sixth day after session one. Each email directed the participant to the study websites and asked them to complete a specific task on each website. Although the notification emails did not explicitly instruct participants to log in, participants needed to do so in order to complete the tasks. In total, the second session required about thirty minutes of the participant's time.

Session Three. The third session took place one week after the initial session. Participants returned to the lab for a thirty minute session. They were asked to complete a final task on each website, and then to fill out the post-test questionnaire (Appendix E). The post-test questionnaire asked participants to rate the usability of the study websites and password

system, as well as about their current password habits. Finally, participants received their payment for the study (whether the $20 honourarium or the 2.0 course credits), and were debriefed (Appendix F).

If a participant could not remember their password in session one, the experimenter reminded the participant of their password, and assisted them in remembering their original password. If they could not remember their password at any later point in the study, they were allowed to reset it. Resetting their password assigned them a new random password (using the appropriate scheme) for that account. If they were in the Chosen Text condition, resetting their password allowed the user to select a new text password.

Hypotheses

As explained above, the independent variable in the study was the password system used, and the five study conditions were Blank Passtiles, Image Passtiles, Object Passtiles, Assigned Text, and Chosen Text passwords. Except for the Chosen Text condition, all passwords were assigned randomly. Based on previous work in the field of graphical passwords and our investigation into the memory literature, three hypotheses were formulated.

Password memorability was measured via two dependent variables: the memory time, and the number of password resets. It was expected that the three graphical password conditions would be more memorable than the Assigned Text password condition, owing to the picture superiority effect. While it was expected that differences between the three graphical password conditions would be seen, the directionality of these differences was unclear since the five password systems have individual characteristics that affect memorability (such as familiarity and the type of memory retrieval harnessed). Since users were allowed to choose their own Chosen Text passwords, and most users were very familiar with this type of passwords, it is expected that they would be easy for participants to remember.

The first hypotheses about memorability refers to the memory time, or the longest duration of time that a participant could be shown to have remembered a password.

H1(a)$_0$: There will be no significant differences in memory time between the Assigned Text condition and any of the three graphical password conditions.

H1(a)$_1$: The memory time will be significantly smaller for the Assigned Text password condition than for the three graphical password conditions.

H1(b)$_0$: There will be no significant differences in memory time among the three graphical password conditions.

H1(b)$_1$: There will be significant differences in memory time among the three graphical password conditions.

H1(c)$_0$: There will be no significant differences in memory time between the Chosen Text condition and any of the other conditions.

H1(c)$_1$: The memory time will be significantly larger for the Chosen Text condition than for any of the other conditions.

The second set of hypotheses about memorability referred to the number of resets.

H2(a)$_0$: There will be no significant differences in the number of password resets between the Assigned Text condition and any of the graphical password conditions.

H2(a)$_1$: There will be significantly more password resets in the Assigned Text password condition than in any of the three graphical password conditions.

H2(b)$_0$: There will be no significant differences in the number of password resets among the three graphical password conditions.

H2(b)$_1$: There will be significant differences in the number of password resets among the three graphical password conditions.

H2(c)$_0$: There will be no significant differences in number of password resets between the Chosen Text condition and any other study condition.

H2(c)$_1$: There will be significantly fewer password resets in the Chosen Text password condition that in any other condition.

As a measure of the usability of the password systems, the time taken to log in was examined. Login time was affected by factors such as familiarity and the type of memory retrieval harnessed, but the physical design of the system and the perceptual tasks needed to complete the login also affected the login times. It was expected that differences would be seen in the login times for the systems, but the directionality of the differences was unclear.

$H3_0$: There will be no significant differences in login times among the five study conditions.

$H3_1$: There will be significant differences in login times among the five study conditions.

Analysis Plan

Memorability

Memory Time. To test H1(a), three *t*-tests were planned on the mean memory memory times for different conditions. *t*-tests were planned between Assigned Text and Image Passtiles, between Assigned Text and Blank Passtiles, and between Assigned Text and Object Passtiles. The *t*-test compares two means that are assumed to come from normal distributions with equal variances where the two samples are assumed to be independent. Normality of the sample was planned to be assessed by examining histograms of the distributions, and examining the skewness and kurtosis of the distribution. Histograms were to be evaluated for evidence of central tendency, and for skewness and kurtosis statistics, 3.00 was used as the limit. If the assumption of normality was not met, a Wilcoxon test (sometimes known as a Mann-Whitney U test) was to be substituted.

To test H1(b), we planned to conduct a one-way ANOVA for the memory times for the four conditions. Analysis of variance (ANOVA) is a procedure used to compare multiple means. It assumes that the samples are independent, that the errors are normally distributed and that the variance is equal between groups. If the assumptions for ANOVA were not met, a Kruskal-Wallis test was to be used. The Kruskal-Wallis test is a non-parametric test that works similarly to a one-way ANOVA. The Kruskal-Wallis tests the equality of medians and assumes no fixed distributions (although it does assume that the group distributions are identical).

To test H1(c) and investigate differences in memory time between the Chosen Text condition and each of the assigned password conditions, we had planned to conduct a set of four one-way *t*-tests, comparing the memory time for each of the assigned password conditions with the memory time for the Chosen Text condition.

Resets. Since it was expected that the number of password resets would be skewed, the use of non-parametric tests was planned.

To test H2(a), three Wilcoxon tests on the number of password resets were planned to be conducted. Similar to H1(a), comparisons were to be conducted between the Assigned Text and Blank Passtiles conditions, the Assigned Text and Image Passtiles conditions, and the Assigned Text and Object Passtiles conditions.

To examine the differences in number of resets between the three graphical password conditions, we planned to conduct a Kruskal-Wallis test. If significance was shown, we planned to follow up with post-hoc Wilcoxon tests.

To look for differences between Chosen Text and the other conditions, a set of four Wilcoxon tests was planned, comparing the number of resets in the Chosen Text condition with the number of password resets in each of the assigned password conditions.

Usability

Login time. To explore the hypothesis that there would be significant differences in login times between the four conditions, a one-way ANOVA was planned. If the distribution of the data was non-normal, Kruskal-Wallis tests were to be used. If the ANOVA showed significance, post-hoc analysis using *t*-tests (or Wilcoxon tests, if the distribution was non-normal) was planned.

Exploratory Analysis

In addition to the analysis outlined above, exploratory analysis was to be conducted into aspects of the data that it was anticipated might be of interest. As part of this exploration, an examination of multiple password interference was planned, which would investigate the extent to which participants confused their different passwords across accounts. As a measure

Questionnaire	Variable	Test
Demographics	Age	Participant demographics
	Expertise	Participant demographics
	Previous study experience	Participant demographics
Post-test	Password recording	Chi-squared test of how many people wrote down their passwords
	Usability perception	Boxplots, Kruskal-Wallis tests to examine differences by condition

of the interference, the number of times that a participant entered a complete password from the incorrect account was to be measured. It was expected that the systems where no cue was provided (i.e., the text passwords and the recall-based graphical passwords) would be more susceptible to multiple password interference than the cued-recall and recognition-based graphical passwords.

As part of the study procedure, data was collected from several questionnaires. The demographics questionnaire asked about the participant's age, gender, occupation and level of computer expertise. The post-test questionnaire asked participants a series of questions evaluating the web sites and the password systems used in the study. It also asked about any previous experience with graphical passwords, and current computer security habits. As part of the exploratory analysis, an examination of data from these questionnaires was planned. The analysis was also to examine how the qualitative usability data collected in the post-test questionnaire compared to the more objective measures of usability.

Amazon's Mechanical Turk (MTurk) is an online crowdsourcing marketplace where voluntary workers are paid to perform "human intelligence tasks". The idea behind MTurk is that some tasks are more efficiently performed by humans than computers, and MTurk attempts to centralize and leverage this intelligence. Recently, MTurk has been proposed as a source of participants for studies in usable security, and several studies (Kittur, Chi, & Suh, 2008; Kelley, 2010) have investigated the possibility. Advantages of MTurk include the easy availability of large numbers of participants, the diversity of the sample (MTurk workers come from all over the world) and the ability to efficiently run large scale studies. Concerns about MTurk include worries that participants are not sufficiently engaged in the task at hand, and

are attempting to "game" the system. Concern has also been raised about the potential for workers to falsify answers (particularly demographic information), and the reliability of the results obtained. As part of the exploratory analysis, a replication of the study was to be run using MTurk. The goal of the replication was to allow investigation into the reliability of data obtained from MTurk by providing directly comparable data.

Several elements of the study were changed in order to be conducted using MTurk. The consent form was presented to the MTurk participants as a webpage. In order to express their consent to be in the study, the participant would have to type in their name and press a button indicating their consent. The initial lab session was modified so that it could be conducted online. The training modules were adapted into instructive webpages explaining how to use the password systems. The procedure followed in the the first session was to be delivered to participants as a series of instructions (with links to webpages) via the website interface. All questionnaires in the study were conducted online using the LimeSurvey software, and MTurk participants were to be provided with links to the appropriate questionnaires. The session two emails remained unchanged. In session three, the procedure was delivered as a series of instructions. The debriefing forms (both interim and final) were presented on a webpage and also emailed to participants. Appendix I outlines the changes to the study for the MTurk replication.

Results

Hypothesis 1: Memory Time

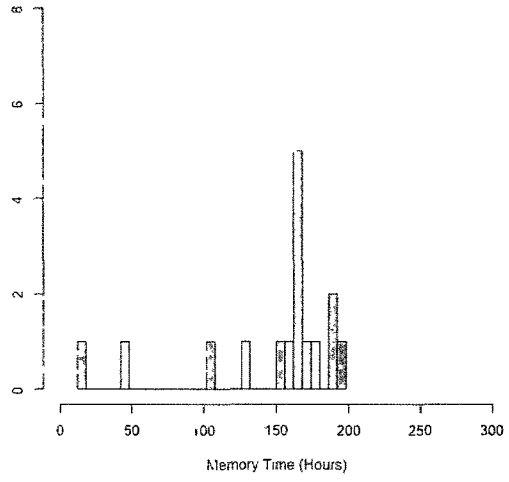
Table 3: Descriptive statistics for memory time (in hours).

	Mean	SD	Median	Skewness	Kurtosis
Blank Passtiles	147.00	50.57	164.14	-1.77	2.57
Image Passtiles	150.76	51.90	167.63	-2.16	4.61
Object Passtiles	160.71	51.96	166.90	-1.63	7.05
Assigned Text	160.67	56.80	166.57	-0.02	5.60
Chosen Text	190.62	44.08	168.73	2.09	3.35

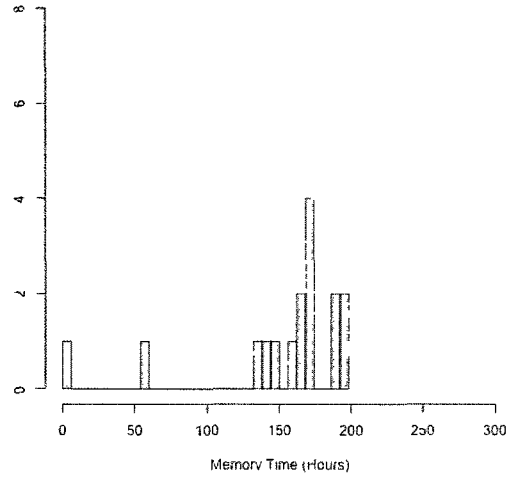
As a measure of the memorability of the password system, each participant's memory time was examined. The memory time was defined as the longest time period between a password creation and a successful login using the same password (i.e., no resets). To aggregate across the three websites used in the study, the mean of each participant's memory time was for each website was taken. One participant was excluded from the memory time analysis because they had to leave town during the study, and completed the last session of the study very late.

Table 3 shows descriptive statistics for the memory time variable. The mean memory time ranged between 147.00 hours in the Blank Passtiles condition and 190.62 hours in the Chosen Text condition. The median memory time ranged between 164.14 hours in the Blank Passtiles condition, and 168.73 in the Chosen Text condition. The total duration of the study was 7 days or approximately 168 hours, so this result showed that most participants were able to remember their passwords for the entire study. Since most participants returned for the second session after exactly 7 days (a few participants returned after 8 or 9 days, due to scheduling constraints), the memory time was limited by this aspect of the study design.

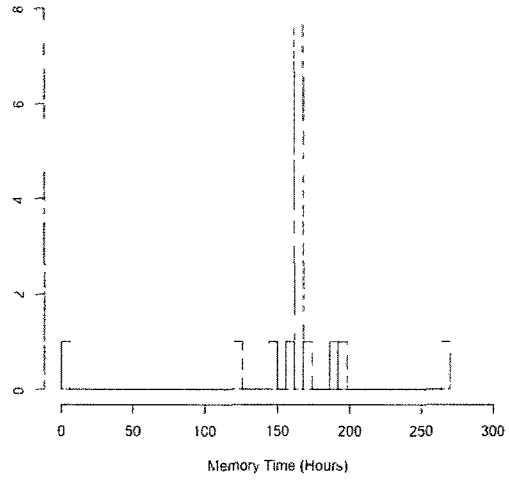
Histograms of the distributions (Figure 7) of memory time (in hours) for each of the five conditions suggested that the distributions are approximately normal. The skewness statistic (Table 3) showed that all the conditions were reasonably centred, but the Image Passtiles, Object Passtiles and Assigned Text conditions were leptokurtotic.



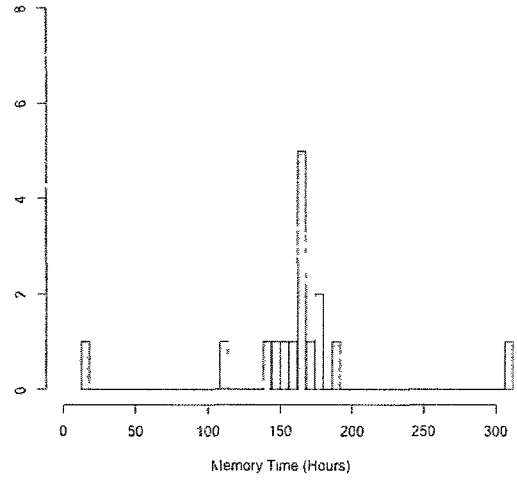
(a) Blank Passtiles



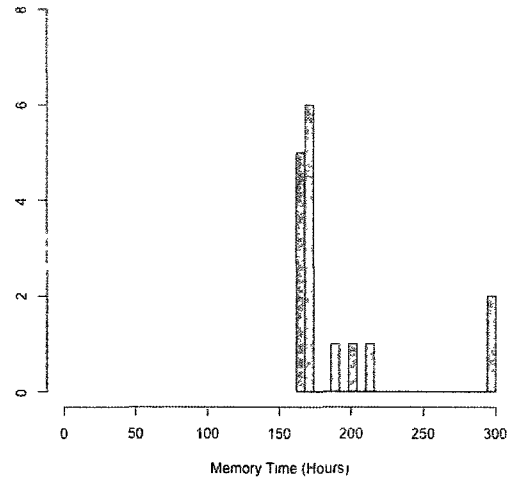
(b) Image Passtiles



(c) Object Passtiles



(d) Assigned Text



(e) Chosen Text

Figure 7. Distributions of memory time for each study condition.

Table 4: t-tests of memory time.

	t	df	p
Assigned Text vs. Blank Passtiles	0.72	30	0.761
Assigned Text vs. Image Passtiles	0.52	30	0.695
Assigned Text vs. Object Passtiles	-0.00	30	0.499

H1(a). To investigate hypothesis H1(a) that the memory time would be significantly smaller for the Assigned Text password condition than for each of the graphical password conditions, a set of three one-sided t-tests (Table 4) was conducted. Each t-test compared a graphical password condition to the Assigned Text condition. No significant differences in memory time between the Assigned Text condition and any of the graphical password conditions was found, indicating that there was no evidence that participants were able to remember graphical passwords significantly longer than Assigned Text passwords.

H1(b). Hypothesis H1(b) said that there would be significant differences in memory time between the three graphical password conditions. To test this hypothesis, a one-way ANOVA of the memory time (Table 5) was conducted. No significant differences in memory time between the three graphical password conditions were seen, indicating that participants were not able to remember their passwords for significantly longer in any of the graphical password conditions.

Table 5: ANOVA comparing memory time for the graphical password conditions.

	df	SS	MS	F	p
PassTiles	2	1607.28	803.64	0.30	0.740
Residuals	45	119253.73	2650.08		

H1(c). To test our hypothesis that the memory time would be significantly longer in the Chosen Text condition than in any of the other conditions, we conducted a set of four t-tests. The t-tests (Table 6) compared the memory time for the Chosen Text condition with the memory time for each of the other study conditions. A significant difference in memory time was seen between Chosen Text and Blank Passtiles ($t(29) = 2.60, p = 0.007$), between

Chosen Text and Image Passtiles ($t(29) = 2.60, p = 0.007$), and between Chosen Text and Object Passtiles ($t(29) = 1.76, p = 0.045$).

Table 6: t-Tests of memory time.

	t	df	p
Chosen Text vs. Blank Passtiles	2.60	29	0.007
Chosen Text vs. Image Passtiles	2.34	29	0.013
Chosen Text vs. Object Passtiles	1.76	29	0.045
Chosen Text vs. Assigned Text	1.67	28	0.053

Hypothesis 2: Resets

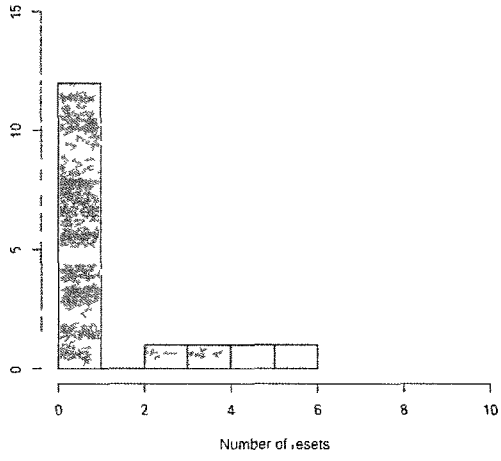
Table 7: Descriptive statistics for password resets.

	Mean	SD	Median	Skewness	Kurtosis
Blank Passtiles	0.88	1.67	0	1.70	1.61
Image Passtiles	0.75	1.73	0	2.55	5.98
Object Passtiles	0.12	0.50	0	4.00	16.00
Assigned Text	0.62	1.15	0	2.07	4.26
Chosen Text	0.00	0.00	0	0.00	-3.66

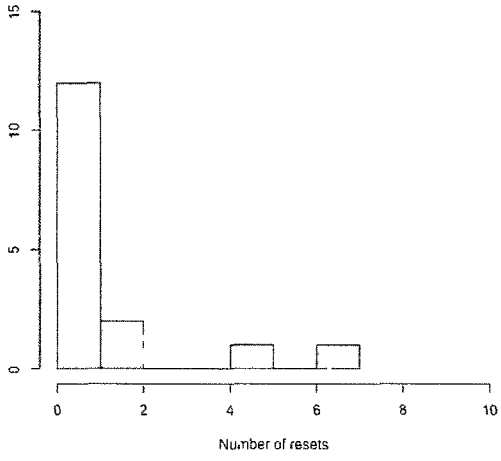
The number of password resets was recorded as a measure of the memorability of the password systems. For each participant, the sum of the number of resets per website was taken. Participants were free to reset their passwords at any time during the at-home sessions of the study.

Participants did not often reset their passwords. The mean number of password resets per condition ranged from 0 in the Chosen Text condition to 0.88 in the Blank Passtiles condition. Of the graphical password conditions, Object Passtiles had the lowest mean number of resets at 0.12 resets per participant. The median number of resets was 0 for all conditions, indicating that most participants never reset any of their passwords. Table 7 shows descriptive statistics for password resets.

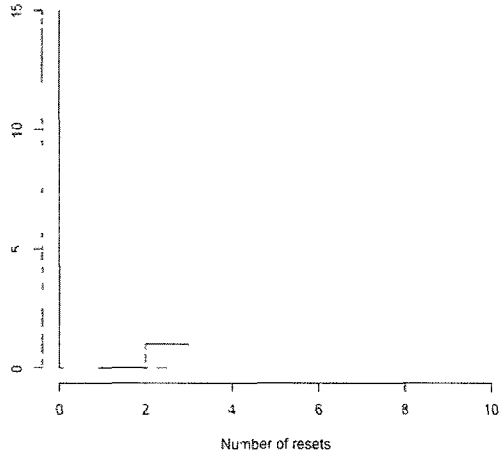
Figure 8 shows the distributions of password resets for each condition. As seen in the histograms, the distributions of resets are skewed and kurtotic, and this is supported by the statistics for skewness and kurtosis (Table 7). As expected, parametric tests were not suitable.



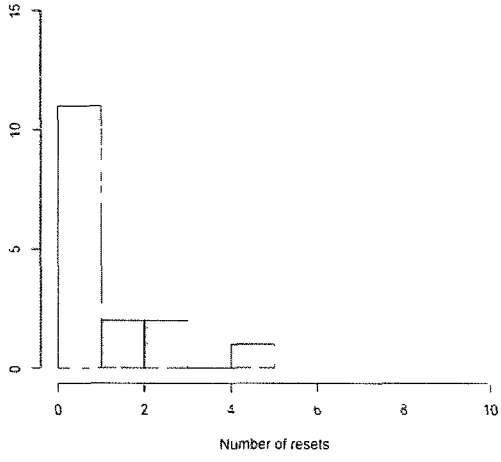
(a) Blank Passtiles



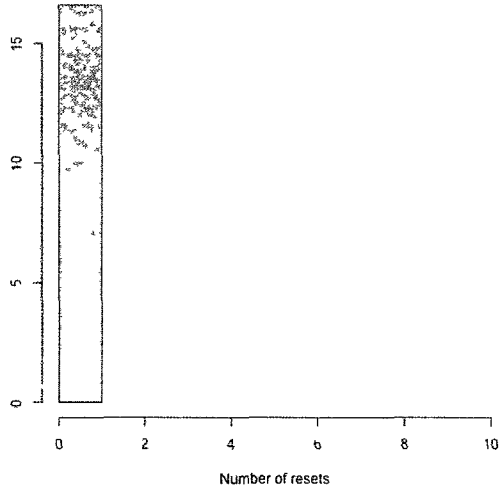
(b) Image Passtiles



(c) Object Passtiles



(d) Assigned Text



(e) Chosen Text

Figure 8. Distributions of resets for each study condition.

In place of t-tests, Wilcoxon tests were conducted, and in place of one-way ANOVAs, Kruskal-Wallis tests were used.

Table 8: Wilcoxon tests of password resets comparing the assigned text condition with each of the graphical password conditions.

	U	p
Assigned Text vs. Blank Passtiles	129.50	0.481
Assigned Text vs. Image Passtiles	134.50	0.388
Assigned Text vs. Object Passtiles	159.50	0.043

H2(a). To test the hypothesis that there would be significantly more password resets in the Assigned Text condition than in any of the three graphical password conditions, a set of three one-way Wilcoxon tests was conducted. Table 8 shows the results of the Wilcoxon tests, each of which compared the Assigned Text condition to one of the graphical password conditions. No significant difference in the number of password resets was seen between Assigned Text and either Image or Blank Passtiles. However, there was a significant difference in the number of password resets between Assigned Text and Object Passtiles ($U = 159.50, p = 0.043$), and participants reset their passwords significantly more often in the Assigned Text condition.

Table 9: Kruskal-Wallis test of resets in the graphical password conditions.

	χ^2	df	p
Kruskal-Wallis chi-squared	2.54	2	0.282

H2(b). Hypothesis 2(b) stated that there would be significant differences in the number of password resets among the three graphical password conditions. To test this hypothesis, a Kruskal-Wallis test (Table 9) was used. No significant differences in the number of password resets among the three graphical password conditions were seen, indicating that there was no evidence that participants reset their passwords differently in any of the Passtiles conditions.

H2(c). It was hypothesized that there would be significantly fewer password resets in the Chosen Text condition than in any of the assigned password conditions. To test this

Table 10: Wilcoxon tests of resets comparing the chosen text condition with each of the other conditions.

	U	p
Chosen Text vs. Blank Passtiles	102.00	0.017
Chosen Text vs. Image Passtiles	102.00	0.017
Chosen Text vs. Object Passtiles	127.50	0.166
Chosen Text vs. Assigned Text	93.50	0.008

hypothesis, a set of four one-way Wilcoxon tests (Table 10) was conducted. The tests showed a significant difference in the number of password resets between Chosen Text and Blank Passtiles ($U = 102.00, p = 0.017$), between Chosen Text and Image Passtiles ($U = 102.00, p = 0.017$), and between Chosen Text and Assigned Text ($U = 93.50, p = 0.008$). No significant difference in the number of password resets was seen between Chosen Text and Object Passtiles.

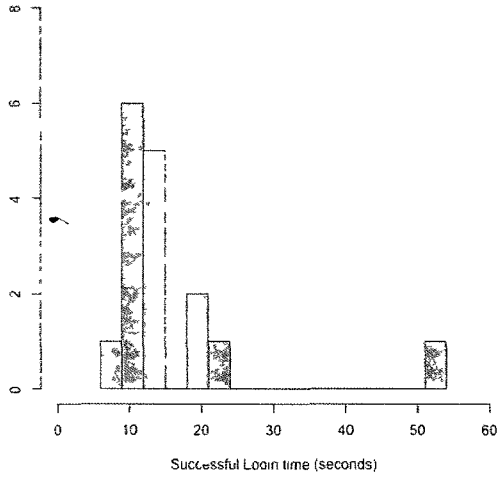
Hypothesis 3: Login times

As a measure of the usability of the password system, the time it took participants to successfully login was measured. In an effort to measure only valid password entry times, only the occasions where users were able to successfully retrieve their passwords were included in the analysis. For each participant, the mean of their successful login times across all three websites was taken.

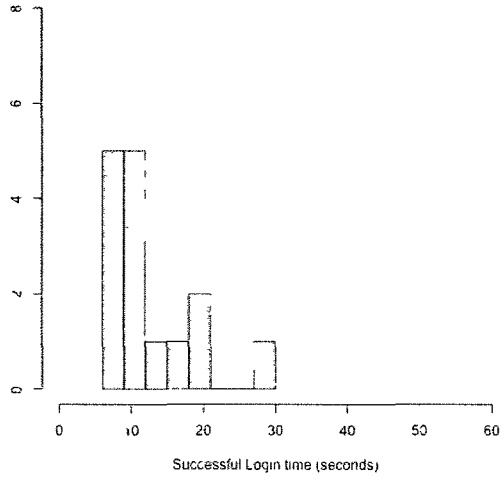
Table 11: Descriptive statistics for login times.

	Mean	SD	Median	Skewness	Kurtosis
Blank Passtiles	15.74	10.48	13.38	3.00	10.19
Image Passtiles	12.65	6.14	9.33	1.38	1.52
Object Passtiles	34.61	21.79	24.25	1.74	2.03
Assigned Text	9.06	3.92	8.28	2.12	5.76
Chosen Text	6.35	2.41	6.00	1.71	3.87

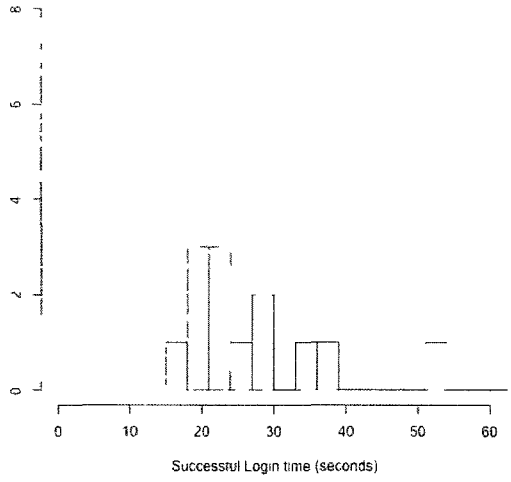
Table 11 shows descriptive statistics for login times. Mean login times were similar and shortest in the Chosen (6.35 seconds) and Assigned Text conditions (9.06 seconds), similar in the Image (12.65 seconds) and Blank Passtiles conditions (15.74 seconds), and longest in the Object Passtiles (34.61 seconds). The median login times were shorter in every condition, indicating that outlying login times were affecting the mean.



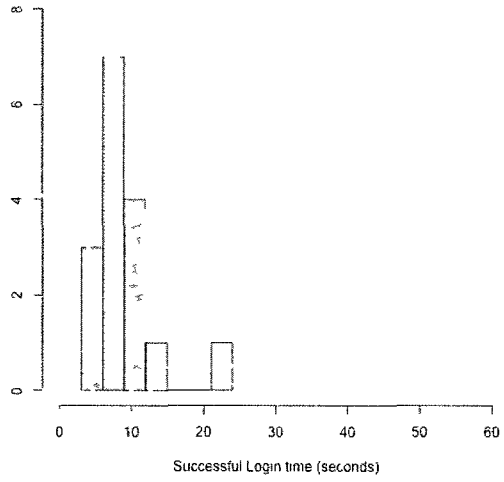
(a) Blank Passtiles



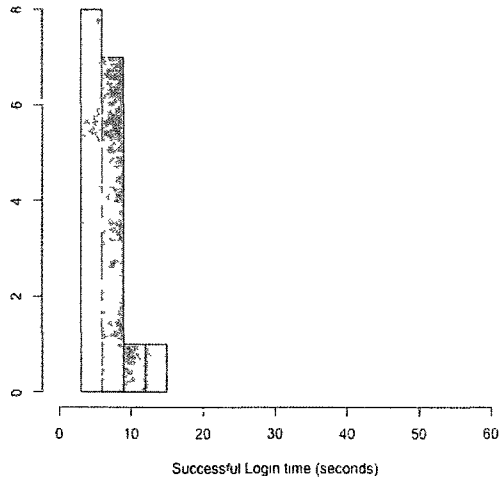
(b) Image Passtiles



(c) Object Passtiles



(d) Assigned Text



(e) Chosen Text

Figure 9. Distributions of login time for each study condition.

Login times were longest in the recognition condition, and longer in the cued-recall condition than in any of the recall conditions, seeming to indicate that login times increased with more recognition tasks.

Table 12: Kruskal-Wallis test of login times.

	χ^2	df	p
Kruskal-Wallis chi-squared	51.31	4	< 0.001

As seen in Figure 9 and Table 11, the distributions of login times were right skewed and leptokurtotic. Since these features made the use of parametric tests inappropriate, Kruskal-Wallis tests were used in place of one-way ANOVA and post-hoc Wilcoxon tests were used in place of t-tests.

Table 13: Pairwise Wilcoxon tests of login times using Bonferroni adjustment.

	U	p
Blank Passtiles vs. Image Passtiles	153.00	1.000
Blank Passtiles vs. Object Passtiles	22.00	0.001
Blank Passtiles vs. Assigned Text	213.00	0.014
Blank Passtiles vs. Chosen Text	260.00	< 0.001
Image Passtiles vs. Object Passtiles	13.00	< 0.001
Image Passtiles vs. Assigned Text	169.50	0.527
Image Passtiles vs. Chosen Text	230.50	0.001
Object Passtiles vs. Assigned Text	236.00	< 0.001
Object Passtiles vs. Chosen Text	255.00	< 0.001
Assigned Text vs. Chosen Text	209.50	0.085

It was hypothesized that there would be significant differences in login times among the five study conditions. To test this, a Kruskal-Wallis test (Table 12) was conducted, and showed significant differences in login times ($\chi^2(4) = 51.31, p < 0.001$). Post-hoc pairwise Wilcoxon tests (Table 13), using a Bonferroni adjustment, showed significant differences in login times between Chosen Text and Blank Passtiles ($U = 260.00, p < 0.001$), between Chosen Text and Image Passtiles ($U = 230.50, p < 0.001$), and between Chosen Text and Object Passtiles ($U = 255.00, p < 0.001$). A significant difference in login time was also seen between Assigned Text and Blank Passtiles ($U = 213.00, p = 0.001$), and between Assigned Text and Object

Passtiles ($U = 236.00, p < 0.001$). As well, a significant difference in login time was seen between Blank Passtiles and Object Passtiles ($U = 22.00, p < 0.001$), and between Image Passtiles and Object Passtiles ($U = 13.00, p < 0.001$). These results showed that it took participants significantly less time to log in using text passwords, and significantly longer to log in using Object Passtiles passwords.

Hypothesis Summary. For memorability, no evidence of differences between the study conditions was found. It was expected that participants with Assigned Text passwords would not be able to remember their passwords as long as participants with graphical passwords, but this was not found to be the case.

In a pilot study of Assigned Text passwords (see Research Study section), participants with Assigned Text passwords coped with the difficulty of remembering their passwords by resetting their passwords at most logins. In response to this, the amount of time associated with resetting passwords was increased, so that participants would be motivated not to reset their passwords unless truly necessary. Marked differences in the number of password resets were not seen between conditions, and it appears that participants were not resetting their passwords as a substitute for remembering them.

When assessing the usability of assigned random passwords, it was expected that there would be differences in login time among the study conditions, affected by both the mode of entry (text- or click-based) and the type of memory retrieval needed. The results of the study indicated that typing was a faster entry mode for participants, and that login times appeared to increase as the more recognition was part of the entry task.

Exploratory Analysis

In the exploratory analysis, we were interested in examining alternative measures of password memorability and usability. These exploratory measures included examining the number of failed password entries, and the number of passwords that a participant successfully remembered from the beginning of the study until the end of the study.

Table 14: ANOVA of memory time for all conditions.

	df	SS	MS	F	p
Password System	4	18788.11	4697.03	1.79	0.140
Residuals	75	196791.88	2623.89		

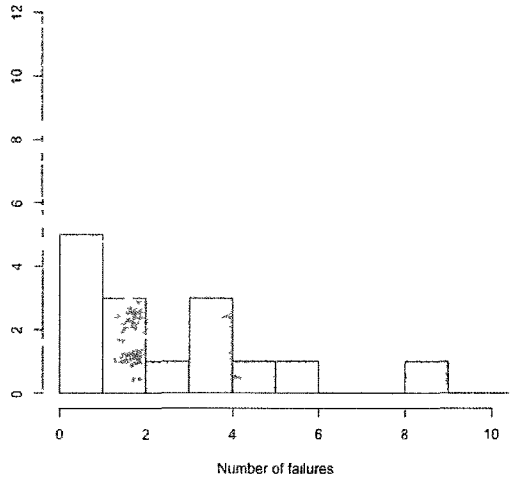
Memory Time ANOVA. To investigate differences in memory time among the study conditions, a one-way ANOVA of memory time (Table 14) was conducted. No significant differences in memory time ($F(4, 75) = 1.79, p = 0.140$) were found among the five study conditions. No evidence of differences could be found between conditions in the length of time that participants were able to remember passwords.

Table 15: Descriptive statistics for failures.

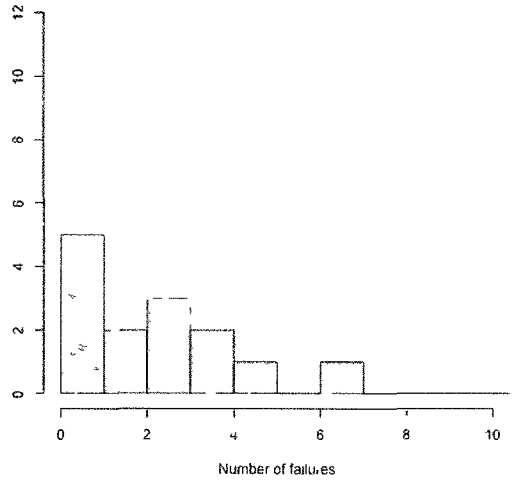
	Mean	SD	Median	Skewness	Kurtosis
Blank Passtiles	2.98	3.26	2.17	1.45	1.88
Image Passtiles	4.00	6.07	2.33	2.67	7.67
Object Passtiles	0.96	1.11	0.67	0.82	-0.76
Assigned Text	1.60	2.39	0.33	1.54	0.79
Chosen Text	0.20	0.29	0.00	0.98	-0.92

Password Entry Failures. As a measure of the memorability of the password system, the number of password entry failures was examined. A password entry failure occurred when a participant incorrectly entered a complete and incorrect password. Because the interest was in password entry attempts that the participant believed to be correct, password failures were not recorded when the participant corrected a password entry attempt without feedback from the system (either by use of the keyboard or by use of the clear button in the graphical password schemes). To explore password entry failures, the sum of failed password entry attempts was calculated on each site and the mean was taken for each participant across the three websites.

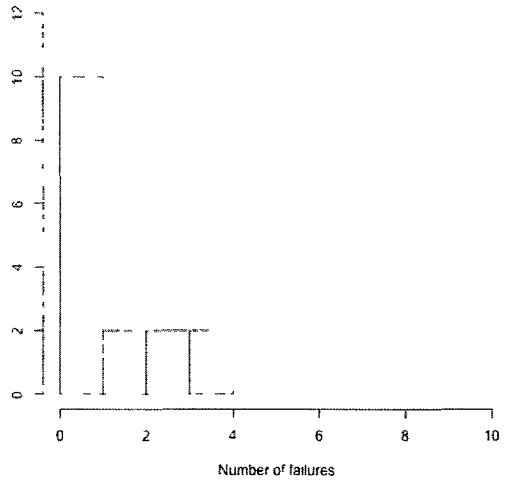
The mean number of password failures for each condition ranged between 0.20 in the Chosen Text condition and 4.00 in the Image Passtiles condition. Although the skewness and kurtosis statistics were not extreme (Table 15), the histograms of entry failures (Figure 10) showed no central tendency and it was concluded that non-parametric tests would be appro-



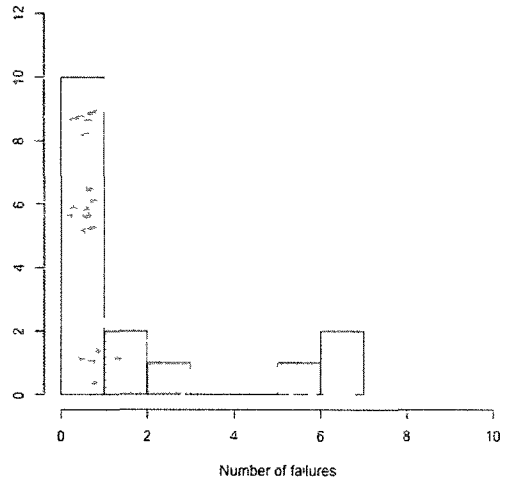
(a) Blank Passtiles



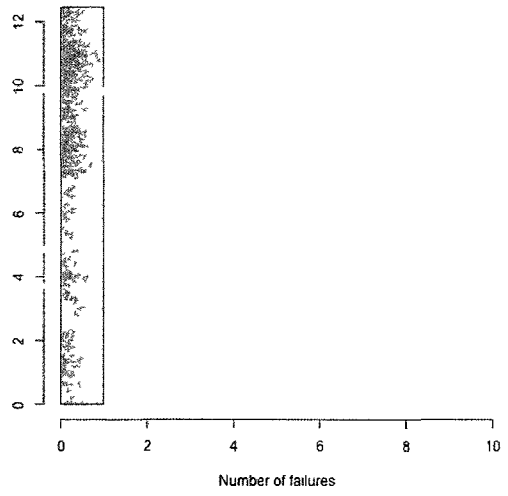
(b) Image Passtiles



(c) Object Passtiles



(d) Assigned Text



(e) Chosen Text

Figure 10. Distributions of number of password entry failures for each condition.

priate.

Table 16: Kruskal-Wallis Test of failed password entries.

	χ^2	df	p
Kruskal-Wallis chi-squared	20.42	4	< 0.001

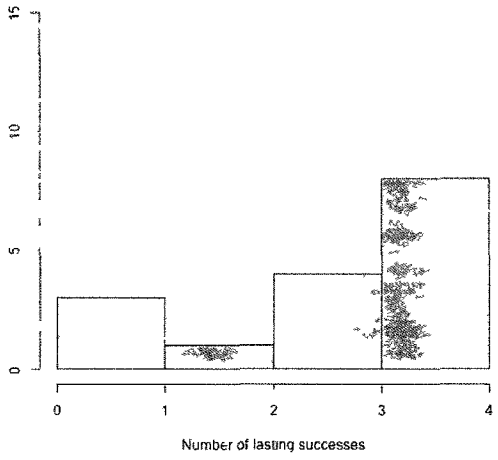
A Kruskal-Wallis test (Table 16) showed a significant effect of password system ($\chi^2(4) = 20.42, p < 0.001$). Post-hoc pairwise Wilcoxon tests (Table 17) using the Bonferroni adjustment showed significant differences in the average number of failures between Chosen Text and Image Passtiles ($p < 0.001$), and between Chosen Text and Blank Passtiles ($p < 0.001$). This indicated that there were significantly more failed password attempts in the Image and Blank Passtiles conditions than in the Chosen Text condition.

Table 17: Pairwise Wilcoxon tests of failed entry attempts using Bonferroni adjustment.

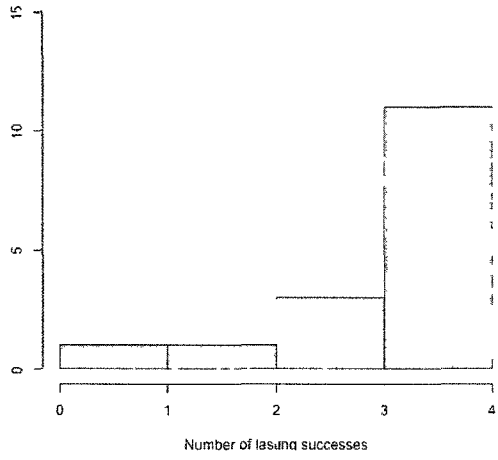
	U	p
Blank Passtiles vs. Image Passtiles	124.00	1.000
Blank Passtiles vs. Object Passtiles	181.50	0.411
Blank Passtiles vs. Assigned Text	162.50	1.000
Blank Passtiles vs. Chosen Text	224.00	0.009
Image Passtiles vs. Object Passtiles	189.00	0.213
Image Passtiles vs. Assigned Text	176.00	0.703
Image Passtiles vs. Chosen Text	243.50	0.001
Object Passtiles vs. Assigned Text	116.50	1.000
Object Passtiles vs. Chosen Text	185.50	0.525
Assigned Text vs. Chosen Text	199.00	0.172

Lasting Password Successes. As another measure of memorability, we were interested in measuring the total number of passwords that each participant successfully remembered throughout the entire study. i.e., At the end of the study, how many of their original passwords did they still remember? The total number of passwords that a participant was able to remember from the beginning of the study (the initial password creation) to the end of the study (the last successful login attempt) was examined. Aggregating across the three websites, this gave a number between 0 and 3 for each participant.

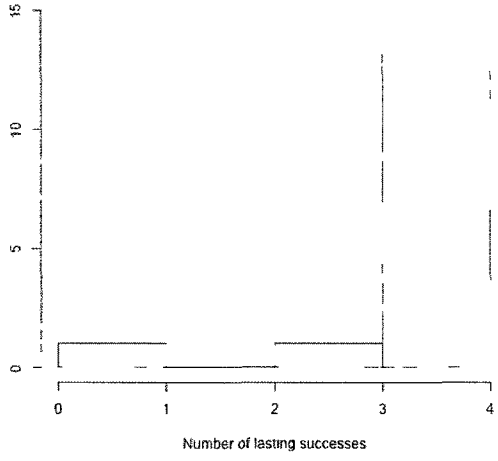
The median number of passwords remembered for the duration of the study was 3 in



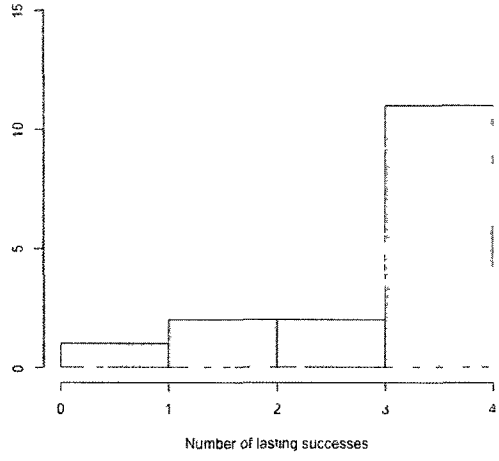
(a) Blank Passtiles



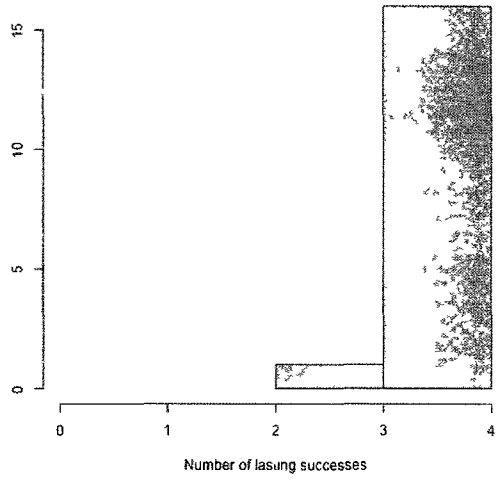
(b) Image Passtiles



(c) Object Passtiles



(d) Assigned Text



(e) Chosen Text

Figure 11. Distributions of number of passwords remembered for the entire study for each study condition

Table 18: Descriptive statistics for lasting successes.

	Mean	SD	Median	Skewness	Kurtosis
Blank Passtiles	2.06	1.18	2.50	-0.97	-0.55
Image Passtiles	2.50	0.89	3.00	-1.92	3.30
Object Passtiles	2.75	0.77	3.00	-3.44	12.22
Assigned Text	2.44	0.96	3.00	-1.59	1.49
Chosen Text	2.94	0.24	3.00	-4.12	17.00

all conditions except Blank Passtiles, indicating that most participants in those conditions were able to remember all of their original passwords for the entire duration of the study. In Blank Passtiles, the median number of passwords remembered for the study duration was 2.5 (Table 18). The distributions of lasting successes can be seen in Figure 11 and along with the skewness and kurtosis statistics, it was concluded that the distributions were left-skewed and non-normal.

Table 19: Kruskal-Wallis test of total number of lasting successes.

	χ^2	df	p
Kruskal-Wallis chi-squared	10.43	4	0.034

A Kruskal-Wallis test of lasting successes (Table 19) showed significant differences ($\chi^2(4) = 10.43, p = 0.034$) between the different conditions. A set of post-hoc Wilcoxon tests using the Bonferroni adjustment showed that the only significant difference in the average number of lasting successes was between the Chosen Text condition and Blank Passtiles condition ($U = 74.00, p = 0.046$).

When the Chosen Text condition was excluded from the analysis, a Kruskal-Wallis test showed no significant differences in the number of lasting successes between any of the assigned password conditions.

Interference. Password interference refers to the possibility that memories of multiple passwords may interfere with each other in memory, causing the user to become confused about which password is associated with which account. Interference causes users to make mistakes when entering their passwords, and can lead to situations where a user enters a password

that is incorrect on the current account, but correct for another account, thereby potentially exposing their other passwords to an attacker.

Table 20: Descriptive statistics for password interference.

	Mean	SD	Median	Skewness	Kurtosis
Blank Passtiles	0.38	0.72	0	1.73	1.70
Image Passtiles	1.06	3.00	0	3.66	13.93
Object Passtiles	0.00	0.00	0	0.00	-3.71
Assigned Text	1.50	3.27	0	3.27	11.66
Chosen Text	0.12	0.49	0	4.12	17.00

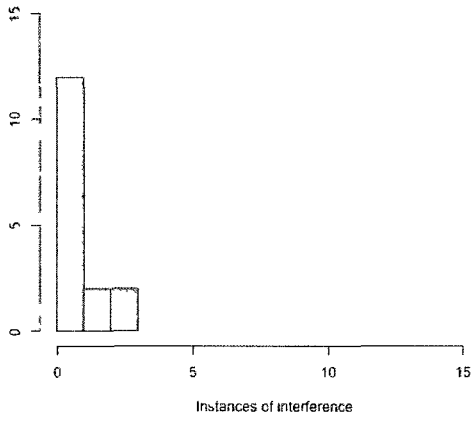
The study had a between-subjects design, and each participant had passwords for three accounts, and the opportunity to experience password interference. Since participants also had the ability to reset their passwords, there existed a possibility for interference with previous passwords for the same account. To measure password interference, the number of times was counted that each participant attempted to log in using a complete password from either another website or any previous passwords from the same website. To aggregate across the three websites, the number of times the participant demonstrated password interference was summed.

Table 21: Kruskal-Wallis test of password interference.

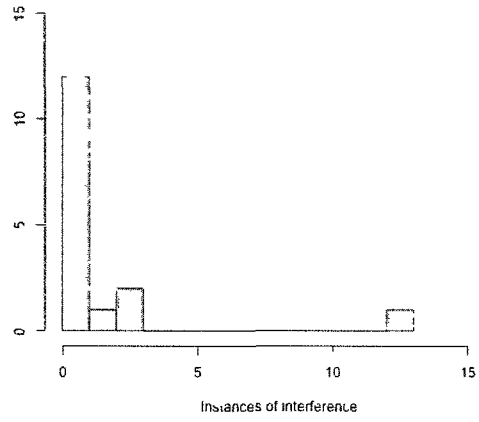
	χ^2	df	p
Kruskal-Wallis chi-squared	10.38	4	0.035

The median number of interference demonstrations was 0 for all conditions (Table 20), but the mean number of interferences was 1.50 for Assigned Text passwords, and 1.06 in the Image Passtiles condition. However, the high mean in the Image Passtiles condition was strongly affected by one participant, who had 12 incidences of interference, 11 of which were on the same account. Figure 12 shows the distributions of interferences for each condition. Since the distributions were leptokurtotic, non-parametric tests were used.

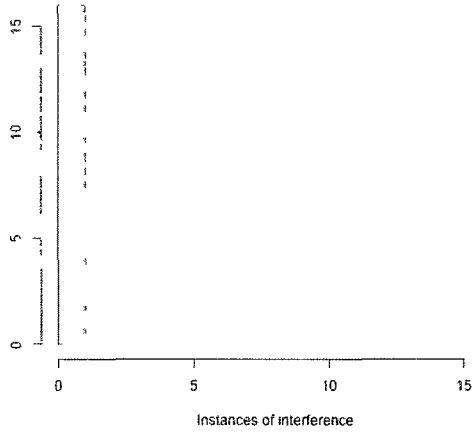
A Kruskal-Wallis test of the number of interference demonstrations was significant ($\chi^2(4) = 10.38, p = 0.035$), but post-hoc pairwise Wilcoxon tests using a Bonferroni adjust-



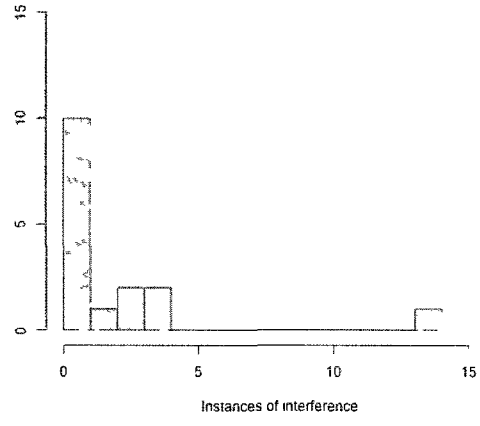
(a) Blank Passtiles



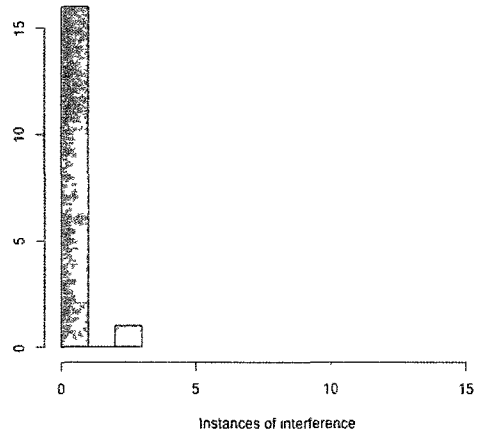
(b) Image Passtiles



(c) Object Passtiles



(d) Assigned Text



(e) Chosen Text

Figure 12. Distributions of instances of interference for each condition.

ment showed no significant differences between any of the conditions. The difference between the instances of interference in the Assigned Text and Object Passtiles conditions was the closest to significant at $U = 176.00, p = 0.086$.

Although password interference was extremely low for the Object Passtiles condition, it was also affected by the lack of overlap in the image sets. The sets of object images used in each account were mutually exclusive, and there was no overlap in images between websites. This made password interference very unlikely, although if a participant had reset their passwords, they could potentially have experienced interference with old passwords from the same website.

Order. One observation made during the in-person sessions was that many participants appeared to be entering their graphical passwords in a consistent order. In the Passtiles passwords, the order that the password tiles were clicked was irrelevant to the success of the password entry attempt (see Research Study section). However, it appeared that some participants were incorporating order as part of the memory task.

Table 22: Descriptive statistics of ordered password entries.

	Mean	SD	Median	Skewness	Kurtosis
Blank Passtiles	3.23	0.89	3.17	-0.55	2.04
Image Passtiles	2.92	1.09	3.00	-0.49	-0.50
Object Passtiles	1.21	0.51	1.00	0.10	2.24

To examine order, the number of times on each website that a participant entered their password in the same order was examined. For each website, the maximum number of times they entered a password in one order was taken. To aggregate across the three websites, the mean was taken across the three website. The analysis for order was limited to the graphical password conditions since entering the correct password components out of order would have led to a password entry failure in the text password schemes. Table 22 shows descriptive statistics for ordered password entries. In the Blank and Image Passtiles conditions, the median number of password entries in the same order was 3, but in the Object Passtiles condition, the median was 1, indicating that most participants did not enter their passwords in a consistent order. The skewness and kurtosis statistics showed that the data had an approximately normal

distribution

Table 23: Pairwise t-tests of order repeats using Bonferroni adjustment.

	t	df	p
Blank Passtiles vs. Image Passtiles	0.89	17	1.000
Blank Passtiles vs. Object Passtiles	1.42	15	0.531
Image Passtiles vs. Object Passtiles	2.09	15	0.163

A one-way ANOVA of order was significant ($F(2, 45) = 25.21, p < 0.001$) but post-hoc t-tests using a Bonferroni adjustment did not show any difference between conditions (Table 23). The most significant difference was in the number of ordered password entries between Image Passtiles and Object Passtiles ($t(15) = 2.09, p = 0.163$).

Since it appeared that participants were entering their graphical passwords in a consistent ordering, we wondered whether that order represented a particular spatial pattern. It was speculated that participants in the Blank and Image Passtiles conditions were choosing an ordering of the password tiles that followed western text conventions: left to right, and top to bottom. In the Object Passtiles condition, because of the shuffling of the object images, ordered password entries would not have led to a consistent geometric pattern. Figures 13 and 14 show heatmap visualizations of the locations in which users clicked on their first, second, third, fourth and fifth password tiles. The heatmaps were created using the SpatStat package (Baddeley & Turner, 2005) and show density functions with radius 1. From the heatmaps, it can be seen that most participants entered their passwords following a top-down, left-to-right order.

To investigate whether the spatial patterns seen in the heatmaps indicated a statistically significant relationship between click number and tile position, regressions of the tile position on the click number of the tile were conducted. To create a measure of tile position and its relation to the top-left corner of the grid, the distance between each click point and the top left corner of the grid was calculated.

Table 24 shows the results of the linear regression for Blank Passtiles. Click number significantly predicted distance from the top left corner ($b = 0.73, t(828) = 16.99, p < 0.001$),

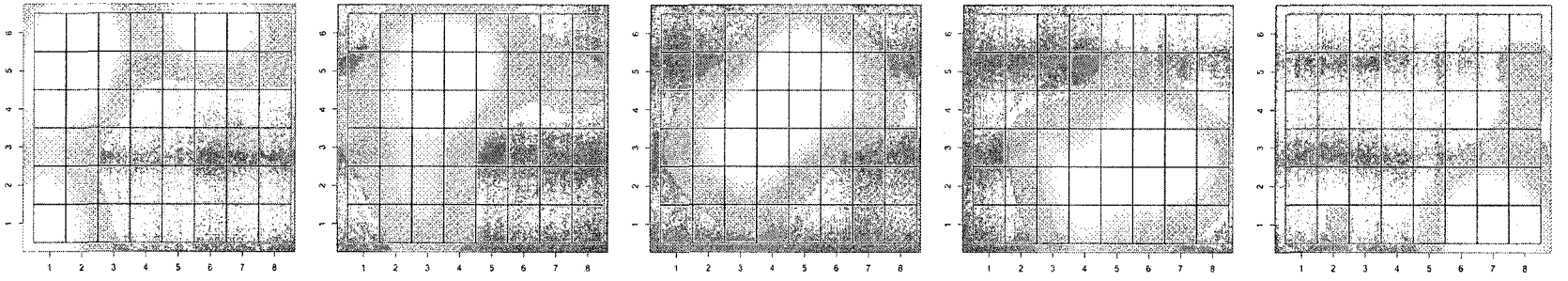


Figure 13. Spatial patterns and directionality of subsequent ordered clicks in the Blank Passtiles condition.

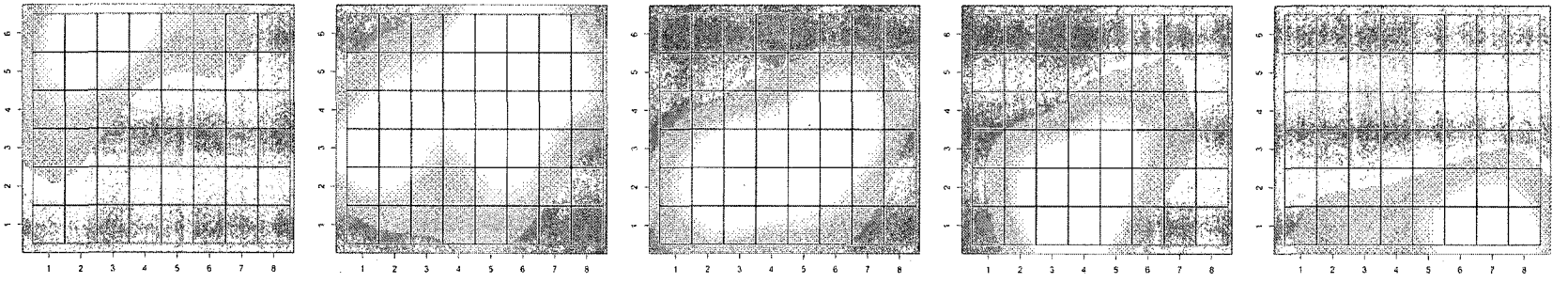


Figure 14. Spatial patterns and directionality of subsequent ordered clicks in the Image Passtiles condition.

showing that participants in the Blank Passtiles condition entered their passwords in a top-down, left-right pattern. Click number also accounted for a significant amount of variance in tile position ($R^2 = 0.26, F(1, 828) = 288.70, p < 0.001$).

Table 24: Linear regression of tile position on click number for Blank Passtiles.

	b	SE	t	p
Intercept	4.04	0.14	28.07	< 0.001
Click number	0.67	0.04	15.33	< 0.001

Table 25 shows the results of the linear regression for Image Passtiles. Again, click number significantly predicted tile position ($b = 0.66, t(803) = 15.20, p < 0.01$), and click number accounted for a significant amount of variance in tile position ($R^2 = 0.22, F(1, 803) = 231.10, p < 0.01$).

The results of the regression analyses show that participants in the Blank Passtiles and Image Passtiles conditions did choose to enter their passwords in a top-down, left-to-right order. This consistency seems to indicate that participants were using a preferred ordering to log in and that this element of user-choice may have been helping users better remember their passwords.

Table 25: Linear regression of tile position on click number for Image Passtiles.

	b	SE	t	p
Intercept	4.19	0.14	30.86	< 0.001
Click number	0.65	0.04	15.92	< 0.001

Password Recording. Although participants were instructed not to write their passwords down, and were restricted from doing so at the time of password creation in the lab, there was no means of controlling their behaviour once they left the lab. In the post-test questionnaire, participants were asked whether they had written down any of their passwords. Since participants were aware that they were not supposed to write down their passwords, it seems likely that our statistics on how many participants recorded their passwords represents a lower bound.

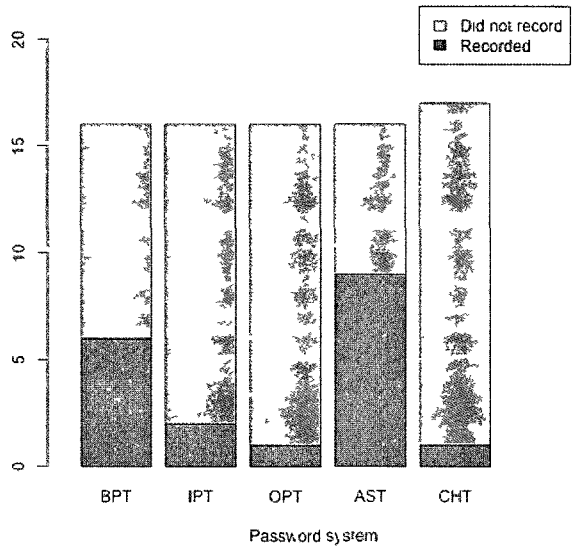


Figure 15. Frequency of reported password recording.

20 participants reported writing down their passwords. The number of participants who recorded their passwords in each condition ranged from 9 people in the Assigned Text condition to just one person in the Chosen Text and Object Passtiles conditions. In the Blank Passtiles condition, 6 people reported writing down their passwords, and 2 people reported writing down their passwords in the Image Passtiles condition. Figure 15 shows a stacked barplot of the frequencies of password recording (and non-recording) in each condition.

Table 26: Chi-squared test of password recordings.

	χ^2	df	p
X-squared	17.97	4	0.001

Since the password recording data were counts, a Chi-squared test was used to look for differences between the four assigned-password conditions. A significant difference ($\chi^2(4) = 17.97, p = 0.001$) in the number of password recordings was seen between the assigned password conditions. Post-hoc pairwise Chi-squared tests using a Bonferroni adjustment showed that the only significant difference in instances of password recording was between the Assigned

Text condition and the Object Passtiles condition ($\chi^2(1) = 7.13, p = 0.046$).

The proportion of participants who reported that they had written down their passwords in the recall conditions was dramatic and likely had a strong effect on the memory time variable, since writing passwords down could allow participants to artificially appear to remember their passwords longer.

Mechanical Turk. A replication of the research study was conducted using Amazon Mechanical Turk as a participant pool. Mechanical Turk has been posited as an easily available source of participants for usable security experiments, but little work has investigated the comparability of results obtained on Mechanical Turk vs. those obtained through more traditional means. A brief overview of the study is presented here, and more detail is given in Appendix J.

77 participants completed the Mechanical Turk (MTurk) study, and the demographics of the MTurk sample were similar to the main study in terms of age and expertise. However, the gender balance was approximately reversed (more men in the MTurk study) and fewer students participated in the the MTurk study. Participants came from around the world, but the largest densities of participants were in the United States (27 participants) and India (26 participants).

The results obtained for memory time were similar to those obtained in the main study. Participants in all conditions were able to remember their passwords for the entire duration of the study. Unlike the main study, it was not found that the mean memory time for Chosen Text passwords was longer than for other study conditions.

Participants in the MTurk study reset their passwords more often than participants in the main study, particularly in the graphical password conditions, which was likely due to the decreased amount of training received by the MTurk participants. A similar pattern to the main study was seen in password resets, where participants reset their passwords significantly less often in the Chosen Text condition than in any of the assigned password conditions.

Similar patterns in login time were seen in both studies, but the login times were longer in the MTurk study, particularly for participants in the Blank and Image Passtiles conditions. The increased login times are likely due to participants with slower computers and the network

delays associated with sending information around the world. Unlike the main study, the login time for Object Passtiles was not significantly longer than for all other conditions, but the login time for Chosen Text was significantly shorter than any other condition.

For more details, statistical tables, figures and some discussion are given in Appendix J.

Security Analysis

In all but the Chosen Text password condition, passwords were randomly assigned, making the effective password space equal to the theoretical password space and pre-determining the security of the password system against guessing attacks. However, in the Chosen Text condition, the password security was largely determined by the participants' choice of passwords.

When creating their passwords, participants were told that their password must be 4 characters long and could have lower case letters and digits. The length of the password was enforced by the system, but the system did not prevent participants from adding characters from different character sets (such as symbols or upper case letters). In order to examine the security of the passwords chosen by participants in this condition, we looked at the character sets used in the passwords, the incidence of password reuse, and at the use of patterns (e.g., dictionary words) in password selection.

The passwords created in the Chosen Text condition included four character sets. They were: lowercase letters, uppercase letters, digits, and symbols. Table 27 shows the frequency of use for these character sets, both alone, and in combination. The most commonly used character sets were lowercase letters and digits: out of 51 passwords created, 57% contained lower case letters and 61% contained digits. Uppercase letters were used by only 2 participants in 6 passwords, and only one instance of a symbol was seen. Combining different character sets is more secure than using only one character set, and several passwords combined multiple character sets. 37% of passwords used more than one character set, and of those passwords, 68% used letters and digits. Only one password was created using characters from three character sets.

Participants were not specifically instructed to use different passwords for their three

Character Set	a-z	A-Z	0-9	Symbol	a-z , A-Z	a-z, 0-9	a-z, A-Z, 0-9
Alone	13	0	18	0	5	13	1
With others	16	6	13	1	1	0	0
Total	29	6	31	1	6	13	1

Table 27: Frequency of character set use in Chosen Text passwords.

different accounts. If a participant asked whether they could reuse passwords, they were told that it would be more secure if they did not repeat their passwords, but the system did not enforce this policy. Of the 17 participants in the Chosen Text condition, 59% had only one unique password, meaning that they reused the same password across all three of their accounts. Reusing passwords across different accounts is a security risk because it creates a single point of failure across multiple accounts and can potentially expose users' passwords to attackers.

A dictionary attack involves an attacker leveraging commonly used patterns to create a prioritized guessing list to use in a more efficient guessing attack. Patterns that can be exploited in text passwords include dictionary words and other common text patterns, such as leading capitalizations, trailing punctuation, and common substitutions, such as '$' for 'S'. In the text passwords created by participants in the Chosen Text condition, we saw examples of all these patterns.

The passwords chosen by users in the study comprise a much smaller effective password space than the assigned random passwords. The US National Institute of Standards and Technology (Burr, Dodson, & Polk, 2006) provides a set of guidelines for estimating the entropy of user-chosen passwords. The relevant parts of these guidelines are quoted below:

> • the entropy of the first character is taken to be 4 bits;
> • the entropy of the next 7 characters are 2 bits per character; this is roughly consistent with Shannon's estimate that "when statistical effects extending over not more than 8 letters are considered the entropy is roughly 2.3 bits per character";
> ...
>
> For user selected PINs the assumption of Table A.1 is that such pins are subjected at least to a rule that prevents selection of all the same digit, or runs of digits (e.g., "1234" or "76543"). This column of Table A.1 is at best a very crude estimate, and experience with password crackers suggests, for example, that users will often

> preferentially select simple number patterns and recent dates, for example their year of birth.

Using the NIST guidelines, the passwords chosen in this study are estimated to have an effective password space of 10 bits (4 bits for the first character, plus 2 bits for each of the 3 succeeding characters). For the passwords consisting entirely of digits, their effective password space is estimated at 9 bits.

The presence of dictionary words and common substitutions further narrows the space. Using the *John the Ripper* password cracking program free dictionaries (Designer, 2011), 8 passwords were able to be guessed using a 9 bit dictionary. A further 4 passwords were able to be guessed using an 11 bit dictionary (that included the smaller 9 bit dictionary). A dictionary of digit combinations was able to guess 18 passwords that included only digits. In total, the John the Ripper dictionaries were able to guess 59% of the created passwords using a dictionary with a size of 13.5 bits.

The effective password space for the Chosen Text passwords was considerably smaller than the password space for the assigned passwords used in the study. Because of this, it is not reasonable to directly compare the memorability or usability of the systems. In the Chosen Text condition, the quantity of information that users are asked to use and remember is far less than in the assigned password conditions. In addition, users are given the opportunity to choose less secure, but more memorable passwords, which is not a choice given to users in the assigned password conditions. Because of the differing security levels, we do not feel the comparison between Chosen Text and the assigned password conditions is fair.

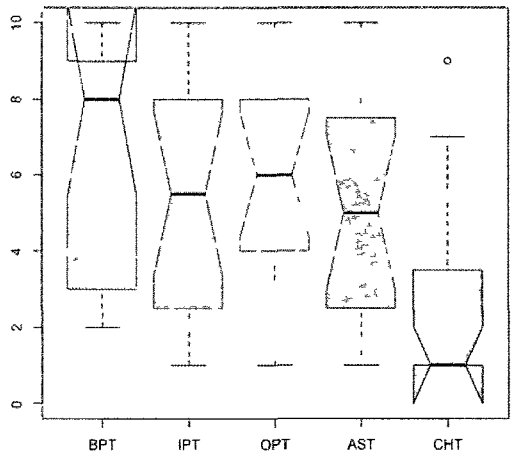
Questionnaire Data

In the post-test questionnaire, participants were asked a series of questions about their perceptions of usability and security in the password system. Participants were given a series of statements and asked to rank their agreement on a scale from 1 to 10, where 1 was "strongly disagree" and 10 was "strongly agree". Since questionnaire data are ordinal data, boxplots were used to examine the distribution of responses to the statements, and Kruskal-Wallis tests were used to look for differences between conditions. Boxplots illustrate the distribution of

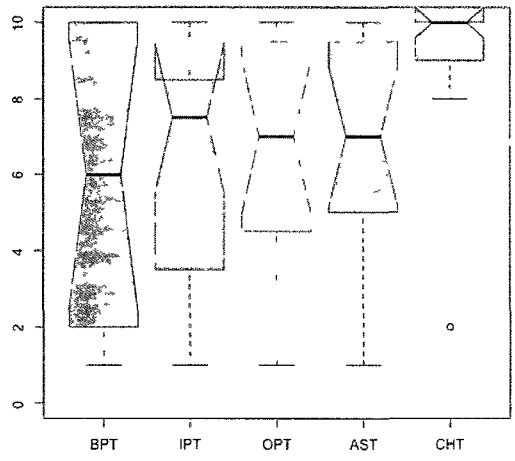
data, showing the median as the black bar in the centre, the 2nd and 3rd quartiles as the bottom and top edges of the "box", and the 1st and 4th quartiles as "whiskers" extending from the box. The notches indicate 95% confidence intervals around the median. In the cases where the confidence intervals fall outside of the 2nd or 3rd quartiles, the notches extend beyond the box.

Participants had significantly more trust in the security of the assigned passwords as demonstrated by significant differences in the responses to the statements "I would trust this password system to protect my financial information" (Figure 16(a)) and "My accounts would be secure if protected by a password like this" (Figure 16(e)). Participants also believed that their assigned passwords would be more difficult for an attacker to guess (Figure 16(f)), and that they would have more trouble guessing other people's assigned passwords (Figure 16(i)).

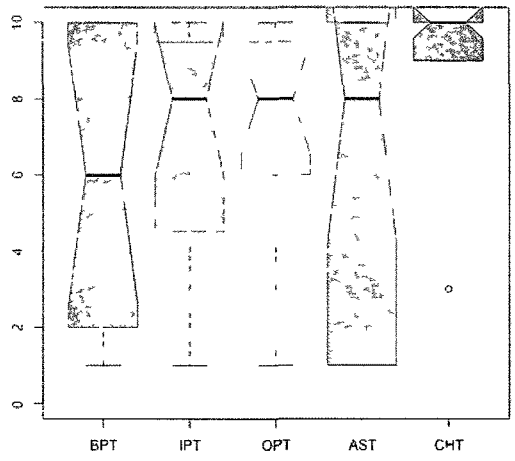
However, participants were less enthusiastic about the usability of the assigned password systems. Participants in the assigned password conditions were less likely to agree that logging in was easy (Figure 16(b)) or accurate (Figure 16(c)). They were also less confident in their ability to remember their passwords over a long period of time (Figure 16(d)). Participants with assigned passwords also felt that logging in was too slow a process (Figures 16(g) and 16(h)).



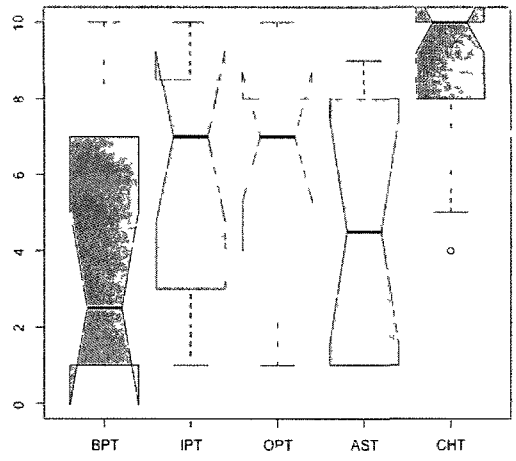
(a) I would trust this password system to protect my financial information



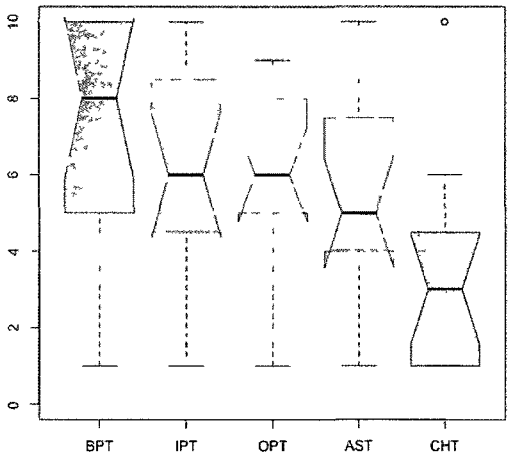
(b) Logging in using these passwords was easy.



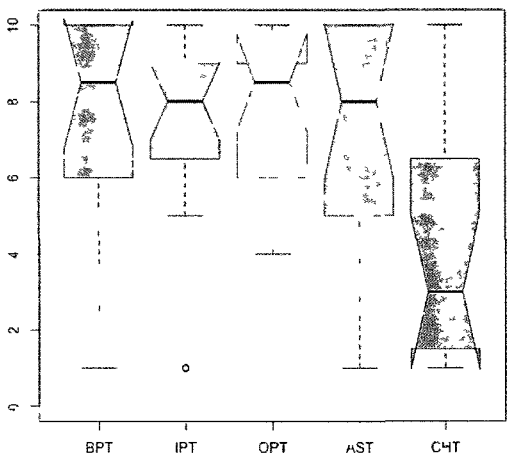
(c) It was easy to accurately enter my passwords.



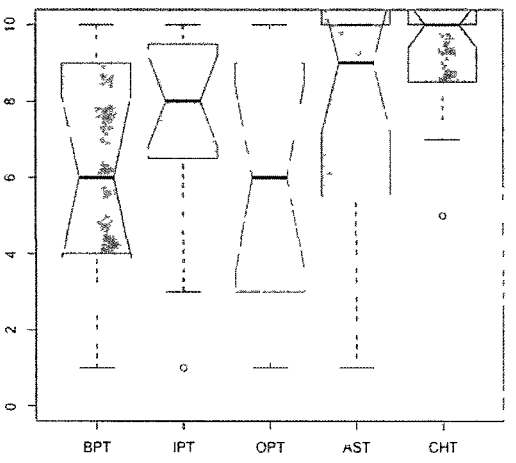
(d) If I didn't log into my accounts for a few weeks, I would still remember my passwords.



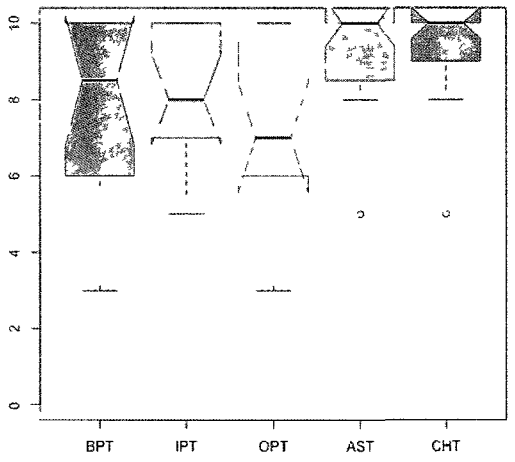
(e) My accounts would be secure if protected by a password like this



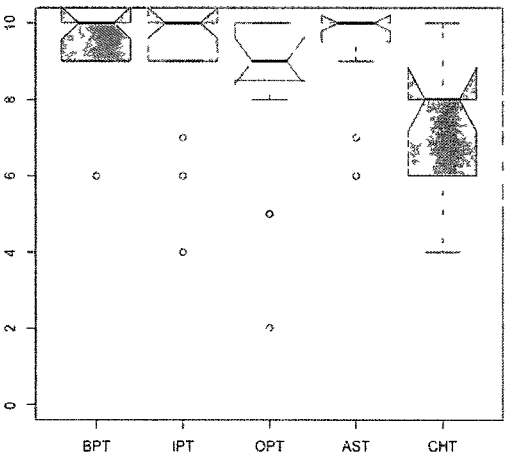
(f) This type of password would be easy for attackers to guess (Graph presented with reversed scale.)



(g) This type of password is too time-consuming. (Graph presented with reversed scale.)



(h) With practice, I could quickly enter my passwords



(i) I believe I could guess other people's passwords. (Graph presented with reversed scale.)

Figure 16. Distributions of Likert scale responses to usability and security perception questions. 1 is strongly disagree and 10 is strongly agree. In 16(f), 16(g) and 16(i) the scales are reversed to show positive attributes consistently.

Discussion

Results Summary

Memorability. The results of the study appeared to show that the passwords were memorable for most users in our study. In all conditions, most participants were able to remember their passwords for the duration of the study, and most participants never reset their passwords. It was expected that participants in the Assigned Text condition would not be able to remember their passwords as long as participants in the graphical password conditions, but this difference was not seen. It was found that participants in the Blank and Image Passtiles conditions had more password entry failures than participants in other conditions and in Blank Passtiles, fewer participants were able to remember all of their original passwords until the end of the study.

The lack of differences in memorability between conditions was a surprise, and seems to point to some potential issues to consider. While participants were instructed not to write down their passwords, and prevented from doing so during the initial password session, many participants reported in the post-test questionnaire that they had written down their passwords during the study, and it seems likely that more people wrote down their passwords than admitted to doing so. The incidence of password recording was higher in the conditions where the password systems relied on pure recall, and the passwords were randomly assigned (i.e., Blank Passtiles and Assigned Text). In the Assigned Text condition, more than half of participants reported writing their passwords down, and approximately 40% of people in the Blank Passtiles condition wrote their passwords down. We speculate that writing their passwords down allowed the participants in these conditions to display increased memory time and fewer resets than they might have otherwise.

Security is most often a secondary task for users (Whitten & Tygar, 1999), and in situations where the security task becomes difficult or burdensome, users develop coping strategies that allow them to bypass the security and focus on their primary tasks. Writing passwords down is a common and simple coping strategy. The high incidence of password recording in the Assigned Text condition seems to suggest that the effort to create a more ecologically valid

situation where security was a secondary task was successful. However, this success can lead to problems in interpreting the results of the study.

In an effort to model security as a secondary task, the study protocol was designed with mildly deceptive elements to focus participants' attention on an artificial primary task. For participants in conditions with novel password systems, the deception may have been obvious, and most participants were unsurprised to learn in the debriefing about our interest in the password systems and various elements of their use. However, for passwords in the Chosen Text condition, the premise of the study as a website usability study seemed plausible, and few participants assumed we were interested in the password systems. This was also true to a certain extent in the Assigned Text condition. The memory time was slightly longer in these conditions, and we speculate that these participants were less aware of the importance of time as a variable, leading them to be less assiduous about completing the sessions on time, and artificially increasing the memory time. In contrast, participants in the graphical password conditions were more aware of true nature of the study, and we speculate that they were making an effort to perform in a way they perceived would assist the study, and completing their tasks on time.

Another possible issue is that we may have underestimated the quantity of information that participants were able to remember. The assigned random passwords used in the study all had a password space of approximately 21 bits, chosen because of a recommendation by Florencio and Herley (2010) that 20 bits of security was suitable for everyday computing environments. However, with the lack of differences in memory time between conditions, it is difficult to know whether the lack of differences stems from a true lack of difference in people's ability to remember 21 bit passwords, or whether it stems from the steps participants took to cope with the study tasks (such as writing passwords down). In a pilot study of assigned text passwords where participants were asked to remember a total of 17 characters in 4 passwords, we found that participants were not able to remember passwords of this length, and resorted extensively to repeated resetting of passwords. Thus, we are hesitant to conclude that the task presented in this study was too easy to show differences between the conclusions.

In an effort to examine an aspect of memorability that might not be so strongly influenced

by password write-downs, we examined the number of password entry failures. Failures were highest in the Image Passtiles condition, with an average of 4 passwords failures per participant. A few participants mentioned that they had been able to remember enough information about their password that they could eventually guess it, but that it had taken them a few tries because their encoded information was not sufficiently specific. For instance, one participant mentioned that she had remembered her password tile was next to a tree, but that she wasn't certain about which side of the tree.

Another measure of memorability was the number of passwords that participants were able to remember from their original creation to the end of the study. It seemed that most participants were able to remember all of their passwords for the duration of the study. Again, it is difficult to assess whether this was because they wrote their passwords down, or because they were genuinely able to remember their passwords.

Usability. To examine the usability of the password schemes used in the study, the time it took participants to successfully login was measured. Since we were interested in the length of time it took participants to enter their correct passwords, only logins where the participants remembered the correct password (i.e., *successful* logins) were included in the analysis. Average login times for the five password systems ranged from fast to slow. The un-cued passwords (Chosen Text, Assigned Text and Blank Passtiles) had median login times of less than 10 seconds, but the median login time was slower for Image Passtiles (approximately 15 seconds) and very slow for Object Passtiles (more than 30 seconds). The login times for Object Passtiles were potentially too slow for use in a real-life situation.

In the graphical password schemes, where the order of tile entry did not affect the correctness of a password entry attempt, we were interested in whether participants were entering their passwords in a consistent order at different logins. The data showed that participants tended to enter their passwords in the same order. In Blank Passtiles and Image Passtiles, where the spatial pattern was consistent across logins, participants tended to choose orderings of their password tiles that followed a top-down, left-to-right pattern. This result suggests that for randomly assigned graphical passwords, allowing logins where the order of entry can

be varied might be an asset to usability. Rather than being forced to remember an arbitrary ordering, users are in effect allowed to pick the most memorable ordering of their random password tiles, without affecting the security of the system. And if users were able to remember an assigned ordering, then a Passtiles system with ordered logins would have a larger password space without increasing the size of the grid, or the number of password tiles in each password.

Password interference occurs when a user enters a password on the incorrect account. Password interference stems from interference in memory, and confusion about which password is associated with which account. Interference leads to password entry failures and poses a risk to the security of password systems when users expose their passwords to other parties as failed login attempts. On occasion, users may be very unsure about which password is exposed to which account, and will enter multiple correct but invalid passwords, potentially giving away all of their passwords to an attacker. In this study, interference was measured by examining the number of times a participant entered a correct password on an incorrect site. Instances of interference were low for most participants, but there was more interference in the Assigned Text condition, exposing a weakness of the scheme for both usability and security.

Writing Passwords Down

In the post-test questionnaire, many participants reported writing their passwords down. Although writing passwords down was a straightforward task for the text passwords, it was less obvious how graphical passwords should be written down. Some participants told us that they drew a grid and marked their gridsquares on the grid. One participant in Blank Passtiles identified and named the shapes created by the tiles in their password (e.g., 'mountain','shoe'), and told the experimenter that they had written down the name of the shapes, but not the actual tile squares. Another participant said that they had written down a string of numbers representing the number of tiles between each of their password tiles in a set ordering. In the Object Passtiles condition, one participant said they had written down the set of object names. In some cases, it seemed that participants were writing down a cue to the memory of the password, but in other cases, they were writing down the password itself.

A few participants mentioned that they had had to write down some of their passwords,

but not all. In the Assigned Text condition, one participant was lucky enough to have one of her passwords be randomly chosen as an English word, and she remarked that she had not had to write that password down to remember it. A participant in the Image Passtiles condition mentioned that the visual elements of one of her background images had been less memorable to her, and she had had to write down the password for that account.

Writing passwords down is not necessarily an enormous risk to security. It does not create a vulnerability to impersonal attacks, where the victim is unknown to the attacker, and can even protect against these attacks by allowing passwords to use more secure passwords. However, writing passwords down can potentially be a risk if the victim is being targeted by someone known to them, and who has access to their office, purse, or home. In spite of this risk, some security experts recommend writing passwords down (Schneier, 2005), since it allows users to remember more passwords, as well as more complex (and thus more secure) passwords. However, experts caution that users should be careful about keeping the list of written passwords in a secure physical environment, such as a locked drawer.

From the perspective of memorability, writing passwords down creates an easy way for users to remember passwords, but introduces a few hindrances to usability. If a user requires the written note to enter the password, there is a potential for copying errors. Also, the user may need to log in from multiple places, necessitating either having several copies of the recorded password, or carrying a copy of the password around. These strategies are arduous for the user, and further increase the potential vulnerabilities of the account. In an ideal world, users would not have to write down their passwords, and could feel confident about remembering all of their passwords at any time.

One goal of the study was to examine the memorability of various kinds of assigned passwords, but to do so in an ecologically valid context that presented security as a secondary task for the user. Usable security literature emphasizes that security is a secondary task for users (Whitten & Tygar, 1999), and this must be accounted for in the design and evaluation of new security systems (Yee, 2004). From the prevalence of password recording in the study, it seems that the goal of modelling security as a secondary task was achieved, but at a cost to the analysis of memorability and usability.

The compensation behaviour seen in this study appears similar to behaviour seen in a study of a persuasive text password system (Forget, Chiasson, van Oorschot, & Biddle, 2008). In the persuasive text password system (PTP), a number of random characters were inserted into user-chosen passwords in an effort to improve the security of the passwords without removing the salience and memorability associated with user-chosen passwords. In the study, participants were asked to create PTP passwords with varying numbers of inserted characters, and it was hypothesized that PTP passwords with more inserted characters would be more secure, but also memorable. However, as the number of inserted characters increased, participants decreased the complexity of their chosen passwords to compensate for the increased memory challenge of having extra characters inserted. The result of this diminished complexity was that the PTP passwords with more inserted characters had similar memorability and security to those with fewer inserted characters. In our study, we saw that as people were asked to remember more complex passwords, they compensated by writing down their passwords in order to perform to a certain standard of memorability.

The effect of password recording on the results of the study suggest that perhaps a different protocol is needed for studying memory retrieval in graphical passwords. Although there would be a heavy cost to the ecological validity, the additional control of a lab-based study could help investigate similar questions without the issues present in the current mixed-location study.

Picture Superiority Effect

The study was designed to investigate the differences in the memorability of assigned passwords between graphical passwords and text passwords. The advantages of graphical passwords are purported to come from the picture superiority effect, which says that people are better at remembering images than words. Since users often choose predictable passwords, regardless of password system, our study focused on the memorability of randomly assigned passwords.

The bulk of the experimental work supporting the picture superiority effect was conducted on the recall and recognition of previously unknown (i.e., random) images. This gave

us reason to believe that there would be usability and memorability advantages for randomly assigned graphical passwords over randomly assigned text passwords.

Owing to the prevalence of password recording in the Assigned Text condition, it is difficult to interpret the results of the study with respect to the picture superiority effect. However, since so many participants in the Assigned Text condition felt the need to write their passwords down, it is possible to interpret this behaviour as an indication that it was harder for participants to remember the assigned text passwords. Significantly more participants wrote their passwords down in the assigned text condition than in the cued-recall or recognition conditions, and we speculate that this was because the assigned text passwords were more difficult for users to remember than those with cues.

However, the format of the assigned text passwords was such that users might have found it very easy to write down their passwords. It might also depend on their behaviour in real life. While many people have established text password-recording habits, few people have any pre-existing behaviours associated with graphical passwords.

Memory Retrieval

Another goal of the study was to explore the impact of different types of memory retrieval (recall, cued-recall, and recognition) on the usability and memorability of randomly assigned graphical passwords. Again, since much of the evidence for the superiority of recognition over recall stems from experiments using random word lists and image sets, we felt there was reason to expect randomly assigned recognition-based graphical passwords to be more memorable than other randomly assigned passwords.

Generate-recognize theory (Anderson & Bower, 1972) says that memory retrieval is a two-step process consisting of the generation of a candidate list of responses, and the recognition of the correct response from the candidate list. Generate-recognize theory explains that recognition is superior to recall because in recognition, the generation step is unnecessary, and the person can apply their effort to recognizing the correct item on the list. While recognition removes the possible errors involved in generating a candidate list, it can still be time-consuming for people to select the appropriate answer from the list.

Encoding specificity theory (Tulving & Thompson, 1973) says that only information encoded at the time of memorization can later be used as cues to memory. Since password entry usually only occurs in one context, this is an advantage for security, since the same information encoded will usually be present at the time of retrieval. In addition, the contextual clues can be a reminder to users not to enter their passwords on the wrong sites.

One finding of the memory literature is that recognition is superior to recall. While it is important that passwords can be accurately retrieved, there are other aspects to password usability that make recognition a less appropriate technique for password usability. The study showed that the time it took users to login in the recognition condition was considerably slower than in the other conditions, and potentially unacceptably slow for real world use.

One facet of recognition memory often unexplored in the memory literature is the time it takes for people to find what they are looking for. This aspect of recognition is unimportant for the accuracy of the retrieval process, but is very important for password systems that leverage recognition memory. Participants often have to enter passwords repeatedly throughout a day, and 30 seconds per login is enough to be disruptive to many users.

In order to ensure that participants were leveraging recognition memory, and not just memorizing the spatial locations of their objects, the Object Passtiles system shuffled the grid of images at every login attempt. Passfaces (Real User Corporation, 2004) is a commercially available recognition-based password system that includes shuffling as protection against shoulder surfing attacks. However, the risk of shoulder surfing attacks is minimal in physically secure locations, and shuffling does not protect against more sophisticated recording attacks, such as those involving a video camera. One way to decrease the login time in an Object Passtiles system might be to skip the shuffle step, and allow users to leverage both recognition memory and their recollection about where the relevant images were located.

Memorization Strategies. Although the Passtiles password systems were set up to encourage users to leverage one specific kind of memory retrieval, it appeared that participants were often leveraging a combination of retrieval techniques.

Participants using the Blank Passtiles system often attempted to translate the pictorial

pattern in front of them into another kind of pattern. Many participants used a technique where they counted the number of tiles, and then memorized the sequence of digits rather than focusing on the visual pattern. This technique was more successful for some participants than others, and seemed to depend on the level of care taken in developing the number string. Some participants were inconsistent in how they counted tiles (counting the number of tiles *between* password tiles, or counting the number of tiles *including* the password tile), introducing an element of uncertainty to the logging in process. Participants sometimes seemed to have difficulty remembering the order of the numbers they had developed, and many participants became confused in situations where their password tiles skipped a row, but they had not encoded that information into the number string. Interestingly, many participants seemed to ignore salient visual details (such as corners or adjacent tiles) in favour of the numerical technique. The level of success with the numerical technique seemed to vary, but it did not appear to be the most effective or efficient strategy for remembering Blank Passtiles passwords. The participants who seemed to have the fewest problems with Blank Passtiles were the participants who took advantage of the visual patterns. One participant referred to the passwords as "spatial passwords" and said her memorization technique had been to imagine her fingers touching the correct tiles, as though playing the piano.

Participants using the Image Passtiles password system seemed to have less trouble learning about how to use the password system than those in the Blank Passtiles condition. Not all participants immediately grasped that the background image could be useful to them, but most figured it out quickly. One participant struggled for a few moments before exclaiming "Oh! I know how I'm going to remember this, that's what the images are for." A few participants used the numerical strategy, but usually only for tiles located in visually indistinct areas of the image.

In the Object Passtiles condition, spatial strategies were ineffective due to the shuffling mechanism, and participants focused largely on the identities of the password objects. Several participants vocalized the names of the objects as they clicked them. A few participants mentioned that they had created mnemonic sentences or scenarios to link the items in their memory. One participant commented that although he could recognize his password images

when presented with the grid, it made him anxious not to have a password he could explicitly recall and rehearse. Since password entry was masked, several participants mentioned difficulties in remembering which tiles they had already clicked. Participants adopted a variety of strategies to cope with this problem, including methodically scanning the grid row-by-row (or sometimes column-by-column, but row-wise seemed more common), or explicitly memorizing the password objects in order, and using that order to log in.

One observation made while watching people learn their graphical passwords was that participants often chose to proceed through the password tiles by row (or occasionally by column). Instead of beginning with the most obvious visual patterns, and learning less obvious tiles in relation to the easily memorable ones, participants chose to go through the squares row-by-row. In Blank Passtiles, this strategy was often combined with the numerical strategy, but participants were seen to be using the strategy in Image Passtiles as well. It seemed that creating a firm and predictable ordering helped users remember their passwords.

The observations made during the study seem to point to the coping strategies that users develop to remember their passwords. In the Blank Passtiles condition, although no cue was given, participants showed a tendency to "translate" the information into a semantic form. In the Object Passtiles condition, participants seemed to want to leverage recall memory alongside recognition memory, and named and repeated their object lists. We speculate that this behaviour was an attempt to scaffold participants' own memory, and provide a recall-able memory that could be leveraged in the case that recognition failed.

Passtiles Evaluation

For the most part, participants in the study had no troubles understanding how Passtiles worked, or what they had to do to log in. Participants often asked for clarification about the instruction that order did not matter, but it appeared this instruction was simply unexpected, and not confusing once confirmed.

The practice mechanism in Passtiles was easily understood by users, and extensively used. Participants particularly liked the "hide" and "show" buttons, and made good use of them in practicing entering their passwords. Although the Passtiles practice system assumes

a secure environment for password creation, it appeared it was worth it to allow users the opportunity to completely memorize their password in the same context of use they would be using for login.

One observation made during the user study was that in a system with assigned passwords and a practice system, the password confirmation step was largely meaningless. Since passwords were assigned to users, the confirm step did not have the same function as in a user-chosen system. Instead of confirming that they had typed the same password, the confirm step existed more as an opportunity for the participant to enter the password they were being asked to remember. However, with the addition of the password practice system, users had already had a chance to enter the password, and felt confident that they knew the password when they clicked "continue" in the password creation window. We found that participants were occasionally confused by the confirm step, and sometimes forgot it. This might suggest that in password systems with integrated practice mechanisms, the separate confirm step is unnecessary.

Conclusion

This study examined the usability and memorability of assigned random graphical passwords. In particular, it investigated how the different forms of memory retrieval affect the memorability of assigned passwords. Assigned text passwords were compared to three different kinds of assigned graphical passwords, each leveraging a different kind of retrieval: recall, cued-recall, or recognition. An additional comparison was made to user-chosen text passwords.

The week-long study had a between-subjects design and took place both in the lab and at home. The results of the study showed that participants appeared to be able to remember all kinds of passwords for the duration of the study, and that most participants never reset their passwords. Usability of the different schemes varied, but in general, recall-based passwords had faster login times than cued-recall or recognition-based passwords. The hypothesis that there would be significant differences in login time among the five conditions was supported by the results of the study, and we found that the login times for Object Passtiles passwords were slowest, and potentially prohibitively slow for real-world use. The hypotheses about

memorability were not supported by the results, but were complicated by the participants who reported writing their passwords down. We speculate that this may have affected the measures of memorability, making them more difficult to interpret. A condition with user-chosen text passwords was included in the study to provide comparisons to a known password form, but the differences in security between the chosen and assigned password conditions makes it difficult to compare the usability and memorability with that of the assigned passwords.

One potentially important issue in the study results was the large proportion of participants in the assigned text and recall-based graphical password conditions that wrote their passwords down. It is possible that the increase in password recording points to participants having more difficulty remembering these types of passwords, but the results are difficult to analyze.

Future work in this area seeks to investigate some of the possible confounds in the study. A controlled lab study could possibly give more insight into the short-term memorability of the passwords, without the confounding primary tasks and password recording strategies. It could also be interesting to conduct a similar study with higher strength passwords, or to conduct a study investigating whether users are able to cope with having an assigned ordering of password tiles.

References

Adams, A., & Sasse, M. (1999, December). Users are not the enemy. *Commuications of the ACM*, *42*(12), 40–46.

Anderson, J. R., & Bower, G. H. (1972, March). Recognition and recall processes in free recall. *Psychological Review*, *79*(2), 97–123.

Baddeley, A., & Turner, R. (2005). An R Package for analyzing spatial point patterns. *Journal of Statistical Software*, *12*(6).

Baranovskiy, D. (2010, Accessed August). *Raphaël JavaScript Library.* JavaScript Library. Available from `http://raphaeljs.com`

Biddle, R., Chiasson, S., & van Oorschot, P. C. (in press). Graphical Passwords: Learning from the First Twelve Years. *ACM Computing Surveys*, *44*(4).

Blonder, G. (1996). *Graphical Password (U.S. Patent)* (No. 5559961).

Bond, M. (2008, March). *Comments on Gridsure Authentication.* (`http://www.cl.cam.ac.uk/~mkb23/research/GridsureComments.pdf`)

Burbeck, S. (1987). *Applications Programming in Smalltalk-80(TM): How to use Model-View-Controller (MVC).* (`http://st-www.cs.uiuc.edu/users/smarch/st-docs/mvc.html`)

Burr, W. E., Dodson, D. F., & Polk, W. T. (2006). *NIST Special Publication 800-63: Electronic Authentication Guideline* (Tech. Rep.). Gaithersburg, USA: NIST: US National Institute of Standards and Technology.

Chiasson, S., Biddle, R., & van Oorschot, P. C. (2007). A Second Look at the Usability of Click-Based Graphical Passwords. In *Proceedings of the 3rd Symposium on Usable Privacy and Security (SOUPS).* ACM.

Chiasson, S., Deschamps, C., Stobert, E., Hlywa, M., Machado Freitas, B., Chan, G., et al. (2010). *The MVP Web-based Authentication Framework* (Tech. Rep. No. TR-10-19). Ottawa, Canada: School of Computer Science, Carleton University.

Chiasson, S., Forget, A., Biddle, R., & van Oorschot, P. C. (2008). Influencing Users Towards Better Passwords: Persuasive Cued Click-Points. In *HCI 2008.* Liverpool, UK: British Computer Society.

Chiasson, S., Forget, A., Biddle, R., & van Oorschot, P. C. (2009). User interface design affects security: Patterns in click-based graphical passwords. *International Journal of Information Security*, *8*, 387–398.

Chiasson, S., Forget, A., Stobert, E., van Oorschot, P. C., & Biddle, R. (2009). Multiple Password

Interference in Text and Click-Based Graphical Passwords. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS).* Chicago, USA.

Chiasson, S., van Oorschot, P. C., & Biddle, R. (2007, September). Graphical password authentication using Cued Click Points. In *European Symposium On Research In Computer Security (ESORICS), LNCS 4734* (pp. 359–374).

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of Processing: A Framework for Memory Research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–184.

Cranor, L. F., & Garfinkel, S. (Eds.). (2005). *Security and usability: Designing secure systems that people can use.* Sebastopol, USA: O'Reilly.

Davis, D., Monrose, F., & Reiter, M. K. (2004). On user choice in graphical password schemes. In *Proceedings of the 13th USENIX Security Symposium.* Berkeley, CA, USA: USENIX Association.

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.

Designer, S. (2011, August). *John the Ripper Password Cracker.* (`http://www.openwall.com/john`)

Dunphy, P., & Yan, J. (2007). Do Background Images Improve "Draw a Secret" Graphical Passwords. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS).* New York, USA: ACM.

Ellis, H. C., & Hunt, R. R. (1989). *Fundamentals of Human Memory and Cognition* (4th ed.). Dubuque, Iowa: Wm. C. Brown Publishers.

Florencio, D., & Herley, C. (2007). A large-scale study of web password habits. In *WWW '07: Proceedings of the 16th Internation Conference on World Wide Web.* New York, USA: ACM.

Florencio, D., & Herley, C. (2010). Where Do Security Policies Come From? In *Proceedings of the 6th Symposium on Usable Privacy and Security (SOUPS).* New York, USA: ACM.

Forget, A., Chiasson, S., van Oorschot, P. C., & Biddle, R. (2008). Improving Text Passwords Through Persuasion. In *Proceedings of the 4h Symposium on Usable Privacy and Security (SOUPS).*

GrIDsure. (2011). *Pattern Based Authentication* (White Paper). GrIDsure Limited.

Haist, F., Shinamura, A. P., & Squire, L. R. (1992). On the Relationship Between Recall and Recognition Memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*, 691–702.

Hlywa, M. (2011). *Do Houses Have Faces?* Unpublished master's thesis, Carleton University.

Hollingworth, H. L. (1913, October). Characteristic Differences between Recall and Recognition. *The American Journal of Psychology*, *24*(4), 532–544.

Jermyn, I., Mayer, A., Monrose, F., Reiter, M. K., & Rubin, A. D. (1999). The design and analysis of graphical passwords. In *Proceedings of the 8th USENIX Security Symposium.* Berkeley, CA, USA: USENIX Association.

Kelley, P. G. (2010). Conducting Usable Privacy & Security Studies with Amazon's Mechanical Turk. In *Proceedings of the 6th Symposium on Usable Privacy and Security (SOUPS).* New York, USA.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *CHI '08: Proceeding of the 26th annual SIGCHI conference on Human Factors in Computing Systems.* New York, USA: ACM.

Komanduri, S., Shay, R., Kelley, P. G., Mazurek, M. L., Bauer, L., Christin, N., et al. (2011). Of Passwords and People: Measuring the Effect of Password-Composition Policies. In *Proceedings of the 29th Conference on Human Factors in Computing Systems (CHI).* New York, USA.

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

Mintzer, M. Z., & Snodgrass, J. G. (1999). The picture superiority effect: Support for the distinctiveness model. *The American Journal of Psychology*, *112*(1), 113–146.

Monrose, F., & Reiter, M. K. (2005). Graphical Passwords. In L. F. Cranor & S. Garfinkel (Eds.), (pp. 161–179). Sebastopol, USA: O'Reilly.

Nelson, D. L., & Reed, V. S. (1976, January). On the Nature of Pictorial Encoding: A Levels-of-Processing Analysis. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(1), 49–57.

Nelson, D. L., Reed, V. S., & McEvoy, C. L. (1977, September). Learning to order pictures and words: A model of sensory and semantic encoding. *Journal of Experimental Psychology: Human Learning and Memory*, *3*(5), 485–497.

Nelson, D. L., Reed, V. S., & Walling, J. R. (1976). Pictorial Superiority Effect. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(5), 523–528.

Paivio, A. (1971). *Imagery and Verbal Processes.* Holt, Rinehart, and Winston.

Paivio, A., Rogers, T., & Smythe, P. C. (1968). Why are pictures easier to recall than words? *Psychonomic Science*, *11*(4), 137–138.

Real User Corporation. (2004, June). *The Science Behind Passfaces* (Tech. Rep.). Real User Corporation.

Schmitz, C. (2011). *LimeSurvey: The open source survey application.* Available from `www.limesurvey.org`

Schneier, B. (2005, June). *Write Down Your Password.* In 'Schneier on Security'. Available from `http://www.schneier.com/blog/archives/2005/06/write_down_your.html`

Shay, R., Komanduri, S., Kelley, P. G., Leon, P. G., Mazurek, M. M., Bauer, L., et al. (2010). Encountering Stronger Password Requirements: User Attitudes and Behaviors. In *Proceedings of the 6th Symposium on Usable Privacy and Security (SOUPS).* New York, USA: ACM.

stock.xchng - the leading free stock photography site. (2011). http://www.sxc.hu.

Suo, X., Zhu, Y., & Owen, G. S. (2005). Graphical Passwords: A Survey. In *21st Annual Computer Security Applications Conference (ACSAC'05).*

Tao, H., & Adams, C. (2008, September). Pass-go: A proposal to improve the usability of graphical passwords. *International Journal of Network Security*, *7*(2), 273–292.

Tulving, E. (1968). When is recall higher than recognition? *Psychonomic Science*, *10*(2), 53–54.

Tulving, E., & Thompson, D. M. (1973). Encoding Specificity and Retrieval Processes in Episodic Memory. *Psychological Review*, *80*(5), 352–373.

van Oorschot, P. C., & Thorpe, J. (2008, January). On Predictive Models and User-Drawn Graphical Passwords. *ACM Transactions on Information and System Security*, *10*(4).

van Oorschot, P. C., & Thorpe, J. (in press). Exploiting predictability in click-based graphical passwords. *Journal of Computer Security.*

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

Watkins, M. J., & Gardiner, J. M. (1979, December). An appreciation of generate-recognize theory of recall. *Journal of Verbal Learning and Verbal Behaviour*, *18*(6), 687–704.

Weber, R. (2006). *The Statistical Security of GrIDsure* (Tech. Rep.). Statistical Laboratory, The University of Cambridge.

Weir, M., Aggarwal, S., Collins, M., & Stern, H. (2010). Testing metrics for password creation policies by attacking large sets of revealed passwords. In *CCS '10: Proceedings of the 17th ACM conference on Computer and communications security.* ACM.

Weldon, M. S., & Roediger, H. L. (1987). Altering retrieval demands reverses the picture superiority effect. *Memory & Cognition*, *15*(4), 269–280.

Weldon, M. S., Roediger, H. L., & Challis, B. H. (1989). The properties of retrieval cues constrain the picture superiority effect. *Memory & Cognition*, *17*(1), 95–105.

Whitten, A., & Tygar, J. D. (1999). Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0. In *Proceedings of the 8th USENIX Security Symposium.* Washington, USA.

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, J.-C., Brodskiy, A., & Memon, N. (2005, July). PassPoints: Design and longitudinal evaluation of a graphical password system. *International Journal of Human-Computer Studies*, *63*(1-2), 102–127.

Yee, K.-P. (2004). Aligning Security and Usability. In *IEEE Security & Privacy*. IEEE Computer Society.

Appendix A

Ethics Application

Purpose:

The purpose of this study is to examine the long-term usability and memorability of random assigned graphical passwords. The study aims to investigate the whether random assigned graphical passwords are more memorable than random assigned text passwords. It also intends to investigate the effects of leveraging different kinds of memory retrieval (recall, cued-recall and recognition) in assigned graphical passwords. We intend to investigate these questions in a three part study where participants create an initial lab session, complete several online logins, and return to the lab for a final session. The independent variable manipulated in the study will be the password system used, and we plan to use four different kinds of passwords, each leveraging a different kind of memory retrieval.

Security is typically a secondary task for a user: they do not log into their e-mail accounts for the experience of entering a password, but rather to accomplish some end goal not related to security. This can make the study of usable security difficult, since drawing user attention to security mechanisms often causes users to behave unusually. In order to provide a more realistic situation, we will provide our participants primary tasks (such as commenting on a blog post) which require the user to log into their account. To achieve this, we are using realistic websites, designed by us and running on our lab servers, that employ our specialized password systems. We feel that focusing the users attention on a primary task other than the authentication task will improve the ecological validity of the experiment.

Participants:

Participants will be recruited from both the university community and the community at large. They will be contacted through the SONA system as well as recruited via posters on campus (see Appendix G for recruitment notices) and by word-of-mouth. We may also recruit from the wider community, using online resources such as Kijiji and Craigslist. Participants will be students, or adults aged 18 and over. We will only accept as participants people who

have regular access to a computer with the internet and who are accustomed to entering a username and password to access secure websites. We will also check to ensure that they can run the web software required for our password systems.

Materials:

The materials for the study include a computer with an internet connection and a web browser (such as Internet Explorer or Firefox) installed, an informed consent form (Appendix B), a demographics questionnaire (Appendix C), a post-test questionnaire (Appendix E), two debriefing forms (Appendices D and F). The websites used will be non-controversial and designed to be familiar and easy to use, and all images used in graphical password systems are chosen to be non-offensive to viewers.

Procedure:

Participants will schedule an in-lab appointment through the SONA system, or through e-mail, if they were not recruited through SONA. The experiment will take place in three sessions. The first session will be in the lab, where participants will be asked to create several passwords and complete the demographics questionnaire. The second session will be online, and participants will be e-mailed reminders to log into their accounts and perform brief tasks on the websites. The third session will be in the lab, and participants will be asked to complete a final task (and login), and fill in the post-test questionnaire before receiving the payment and debriefing.

Session 1. After being escorted to the experiment room, participants will be given a consent form (Appendix B) to read before the study begins and sign if they agree to participate. They will be told that the purpose of the study is to evaluate the usability of the study websites by performing typical online tasks. Participants will be told that they will be paid their choice of $20 or course credit for their time (if eligible). Participants will be reminded not to reveal personal information, such as real passwords, during the study. Participants will be provided all of the information needed to register for the websites so that no personal information is required or recorded. However, we will need to send e-mail to participants real e-mail addresses.

Participants will first complete a brief training module, instructing them in how to use the password system used in the study. Participants will have the opportunity to practice entering passwords before moving onto the password creation phase. First they will be given an overview of the websites used in the study and the tasks they will be expected to perform (tasks will be simple and easy to complete, and will include activities such as commenting on blog posts or voting in online polls). Participants will use an anonymous username to create an account on each website (with a password). Each password creation will consist of two stages: password selection (where the user first chooses their password) and password confirmation (where the user repeats the chosen password, to emulate real-world password creation). After the passwords have been created, participants will be asked to fill out a demographics questionnaire (Appendix C) asking about their age, gender, occupation, and level of familiarity with internet use. After completing the questionnaire, participants will be asked to log back into the accounts previously created and perform the tasks associated with each account (see Appendix G for a screenshot of the study websites). Participants will be allowed to reset their passwords if necessary. Before leaving, participants will be asked to enter their password a final time to confirm that they remember it at the end of the session.

Session 2. During the following week, participants will periodically be asked by e-mail to log in via the Internet to the accounts they created in the lab. Each participant will receive three emails.

Session 3. The final session will be scheduled two weeks after the initial session. In this session, the participant will return to the lab and be asked to complete a task on all three websites. After they finish, they will be asked to complete the post-test questionnaire (Appendix E), asking participants to rate the usability of the study websites and password system, as well as about their current password habits. Finally, participants will be debriefed and paid.

This experiment poses no more risk to users than ordinary computer use. Participants will be debriefed in two steps: They will be given an interim debriefing form after the in-lab session (Appendix D) and then will be provided a more thorough debriefing form (Appendix F)

after having completed the post-test questionnaire. The latter debriefing form will contain a detailed explanation of the study, as well as contact information.

Deception

We are interested in the experience of using websites that require authentication, including the process of creating passwords and entering them. In order to investigate this subject, we are focusing the participants attention not only on the password system used, but also the websites where they are used. Although we do not draw the participants attention to the password systems, we do tell participants that we are interested in the password systems as part of the overall picture of website usability. This is because we wish to obtain a realistic understanding of the way participants use passwords in the context of everyday web usage. To protect participants, we provide a detailed debriefing, and explain in it why we chose to focus their attention primarily on the websites. We will also inform participants that they may choose to withdraw their data from the study.

Appendix B

Informed Consent

The purpose of an informed consent is to ensure that you understand the purpose of the study and the nature of your involvement. This informed consent must provide sufficient information such that you have the opportunity to determine whether you wish to participate in the study. This study has received clearance by the Carleton University Psychology Research Ethics Board (11-100).

Research Personnel:

Elizabeth Stobert	Dr. Robert Biddle
Principal Investigator	Faculty Sponsor
Carleton University	Carleton University
1125 Colonel By Drive	1125 Colonel By Drive
Ottawa, ON, Canada	Ottawa, ON, Canada
(613) 520-2600 x 1987	(613) 520-2600 x 1987
estobert@connect.carleton.ca	robert_biddle@carleton.ca

Purpose: The purpose of this usability test is to evaluate the usability of several web systems. In particular, we are investigating how easy or hard it is to log in and perform various tasks on several websites.

Task Requirements: We will ask you to log in and complete tasks on various websites. We will then ask you to fill out several questionnaires. At the conclusion of the study, you will be asked to provide us with any suggestions you might have to improve the websites.

Duration and Locale: The first session should take approximately 1 hour, the subsequent online sessions should take approximately 40 minutes in total, and the final lab session will take approximately 20 minutes. The total duration of the study (including the online sessions) will be 2 weeks. You will be paid a $20 honourarium OR receive 2.0 course credits for your time. You will receive your honourarium after completion of the online sessions. The lab session will primarily take place at the HotSoft lab located in room 2110 in the HCI building at Carleton University. If you choose to receive the $20 honourarium, you will be asked to

return to the HotSoft lab to pick up your honourarium.

Potential Risk/Discomfort: There will be no psychological or physical risk, beyond any risk normally involved in using computers.

Anonymity/Confidentiality: All collected data will be held completely confidential. The data will only be made available to those people involved with this study. Data will be coded for identification purposes.

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 also have the right to refuse to answer any questions you do not feel comfortable answering, for any reason and without explanation. You will receive the payment for the completed portions of the study even if you choose to withdraw from the study. If you choose to stop responding to our emails, we will attempt to contact you but if we cannot contact you, we will consider you as withdrawn from the study, and will send you information on how to collect your compensation. To withdraw, please email the study organizers at carletonHCIstudy@gmail.com.

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

Dr. Janet Mantler	Dr. Monique Sénéchal
Chair, Department of Psychology	Chair, Carleton University Ethics Committee for Psychological Research
Carleton University	Carleton University
1125 Colonel By Drive	1125 Colonel By Drive
Ottawa, ON, Canada	Ottawa, ON, Canada
(613) 520-2600 x 4173	(613) 520-2600 x 1155
psychchair@carleton.ca	monique_senechal@carleton.ca

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

Participant's Name:

Participant's Signature:

Researcher's Name:

Researcher's Signature:

Date:

Appendix C

Demographics Questionnaire

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

1. Age:

2. Sex: (M/F)

3. What is your occupation?

a) If you are a student, at what level are you studying? (College, University undergraduate, Masters, Ph.D., Other)

b) If you are a student, what is your field of study? (i.e. academic program)

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

5. How often do you browse the web?

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

6. How often do you check your e-mail?

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

7. What is a blog (weblog)?

8. How often do you comment on blog posts?

- Never
- Daily
- Weekly
- Monthly

- Yearly

9. Have you ever been in a password study before? If so, please describe the study.

10. Have you ever been in a website usability study before? If so, please describe the study.

11. Have you ever used a graphical password (using a picture to enter a password instead of typing in letters and numbers)? If so, please describe it as best you can, and tell us where you used it.

Appendix D

Interim Debriefing

The purpose of this experiment is to study users accessing online systems to complete tasks. We are exploring various methods of authentication and performing typical website tasks. We are evaluating how easily users can log in and perform these tasks using the online systems. Our aim is to improve computer systems to make them more secure and easy-to-use.

More details about this study will be made available once you have finished the entire study.

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

Elizabeth Stobert	Dr. Robert Biddle
Principal Investigator	Faculty Sponsor
Carleton University	Carleton University
1125 Colonel By Drive	1125 Colonel By Drive
Ottawa, ON, Canada	Ottawa, ON, Canada
(613) 520-2600 x 1987	(613) 520-2600 x 1987
estobert@connect.carleton.ca	robert_biddle@carleton.ca

Should you have any ethical concerns about this study please contact, Dr. Monique Sénéchal. Should you have any other concerns please contact Dr. Janet Mantler.

Dr. Janet Mantler	Dr. Monique Sénéchal
Chair, Department of Psychology	Chair, Carleton University Ethics Committee for Psychological Research
Carleton University	Carleton University
1125 Colonel By Drive	1125 Colonel By Drive
Ottawa, ON, Canada	Ottawa, ON, Canada
(613) 520-2600 x 4173	(613) 520-2600 x 1155
psychchair@carleton.ca	monique_senechal@carleton.ca

If you do not receive an email from us in the *next 24 hours*, please email us at: CarletonHCIstudy@gmail.com

If you have any other questions, or if any of your email information changes, please contact us.

Please keep this form to remember your username and next appointment.

Username:

Next Appointment:

Appendix E

Post-test Questionnaire

Please rate the following statements on a scale of 1 to 10, where 1 is strongly disagree and 10 is strongly agree.

1. It was difficult to navigate through the blogs.
2. It was easy to complete my tasks on the blogs.
3. On the blog, the fonts were difficult to read.
4. The layout of the blog website was intuitive.
5. I was able to complete my tasks quickly on the blogs.
6. On the blog website, the wording was hard to understand.
7. The blog website was easy to use.
8. It was easy to create my password.
9. I would trust this password system to protect my financial information.
10. Logging in using these passwords was easy.
11. It was easy to accurately enter my password.
12. I would use a password of this type.
13. If I didnt log in to my accounts for a few weeks, I would still remember my passwords.
14. My accounts would be secure if protected by a password of this type.
15. This type of password would be easy for attackers to guess.
16. My passwords are unlikely to have any meaning to other people.
17. It was difficult to enter my password even though I thought I remembered it.
18. This type of password is too time-consuming.
19. With practice, I could quickly enter my passwords.
20. I believe I could guess other peoples passwords.
21. Text passwords are more secure than this type of password.
22. These passwords are quicker to use than text passwords.
23. Given the choice between a text password and a password of this type, I would choose a password of this type.

24. If I were in a hurry, I would rather enter a text password than this type of password.

25. I prefer text passwords to this type of passwords.

26. I would be happy if computer systems used this type of passwords instead of text passwords.

27. Logging on using a password of this type was easier than with a text password.

28. It would be easier to remember 5 different text passwords than 5 different passwords of this type.

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

30. Do you sometimes re-use the same password on different web sites? (Y/N)

31. How do you decide if a web site is secure?

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

- It is easy for you to remember
- It is suggested by the system
- It is difficult for others to guess
- It is the same as another password you currently have
- Other (please specify):

33. On a scale of 1 (not at all concerned) to 10 (very concerned), how concerned are you about the security of your passwords?

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

Appendix F

Debriefing

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

The usability test was designed to identify problems with using these programs. In this study, we are studying how different kinds of password systems can make it easier or more difficult to remember assigned passwords. We are interested in how the password system used, the number of passwords created, the frequency of login, and the number of logins the user has to enter affects the memorability and usability of the passwords. We hypothesize that graphical passwords will make it easier for users to remember randomly assigned passwords, which have security advantages.

We are interested in the usability of the passwords in the context of everyday use, and for this reason, we chose to draw your focus to the websites rather than the password system. This was needed to collect realistic data about password use. If you wish to withdraw your data from the study, please indicate to the experimenter that you wish to do so.

The results of the usability test will be used to evaluate the practicality of the password systems and to make recommendations on how they can be improved. Your thoughts, comments, and opinions will be taken into consideration in making design recommendations. Thank you for participating in this usability study. Your time and efforts are greatly appreciated!

To learn more about graphical passwords, see:

Graphical Passwords: Learning from the First Twelve Years

R. Biddle, S. Chiasson, P.C. van Oorschot

ACM Computing Surveys 44:4

`http://hotsoft.carleton.ca/~sonia/content/Chiasson_TR-11-01.pdf`

To learn more about memory retrieval, see:

Memory

A. Baddeley, M. W. Eysenck, M.C. Anderson

Taylor & Francis Group: Psychology Press

This study has received clearance by the Carleton University Psychology Research Ethics Board (11-100). If you have any further questions regarding this research, please contact:

Elizabeth Stobert	Dr. Robert Biddle
Principal Investigator	Faculty Sponsor
Carleton University	Carleton University
1125 Colonel By Drive	1125 Colonel By Drive
Ottawa, ON, Canada	Ottawa, ON, Canada
(613) 520-2600 x 1987	(613) 520-2600 x 1987
estobert@connect.carleton.ca	robert_biddle@carleton.ca

Should you have any ethical concerns about this study please contact, Dr. Monique Sénéchal. Should you have any other concerns please contact Dr. Janet Mantler.

Dr. Janet Mantler	Dr. Monique Sénéchal
Chair, Department of Psychology	Chair, Carleton University Ethics Committee for Psychological Research
Carleton University	Carleton University
1125 Colonel By Drive	1125 Colonel By Drive
Ottawa, ON, Canada	Ottawa, ON, Canada
(613) 520-2600 x 4173	(613) 520-2600 x 1155
psychchair@carleton.ca	monique_senechal@carleton.ca

Appendix G

Study Websites

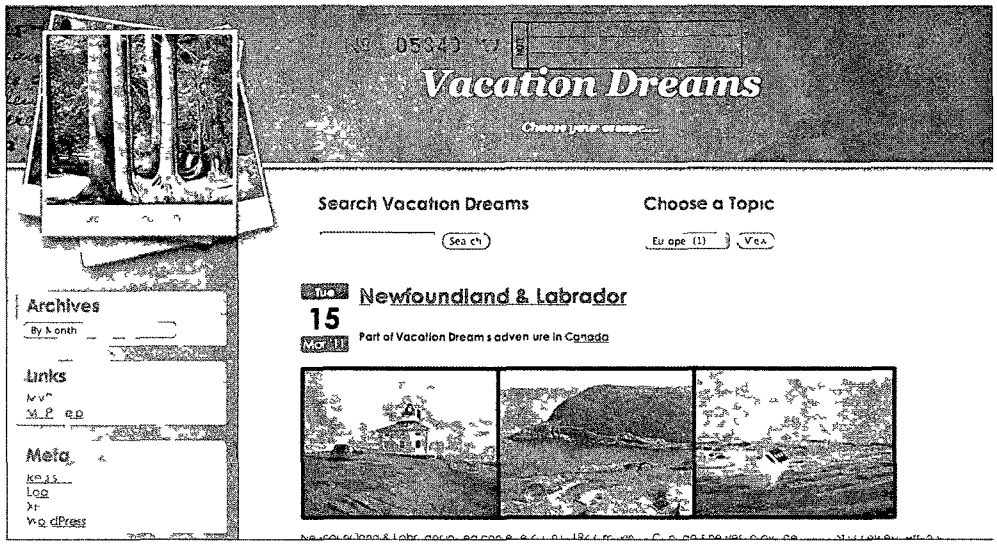


Figure G1 The Vacation Dreams blog

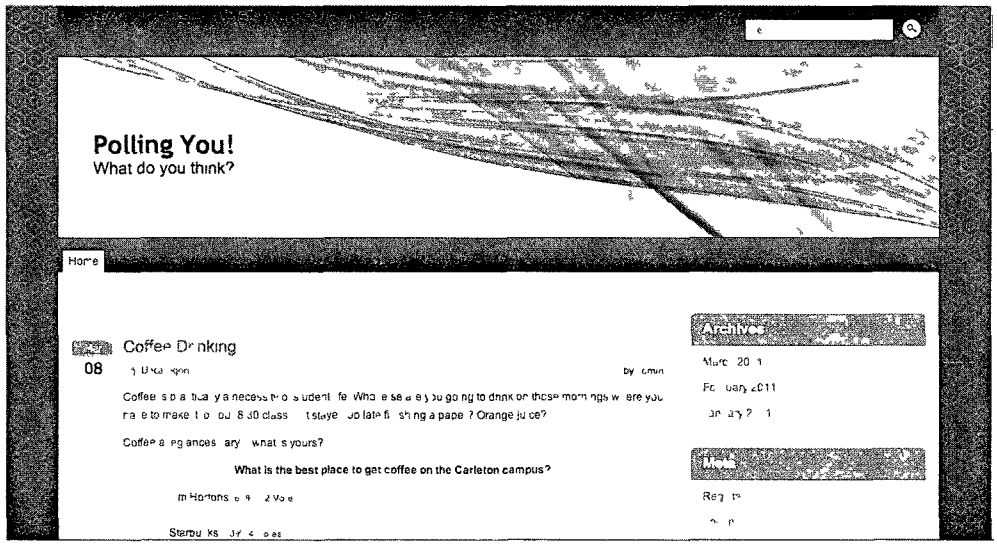


Figure G2 The Polling You! blog

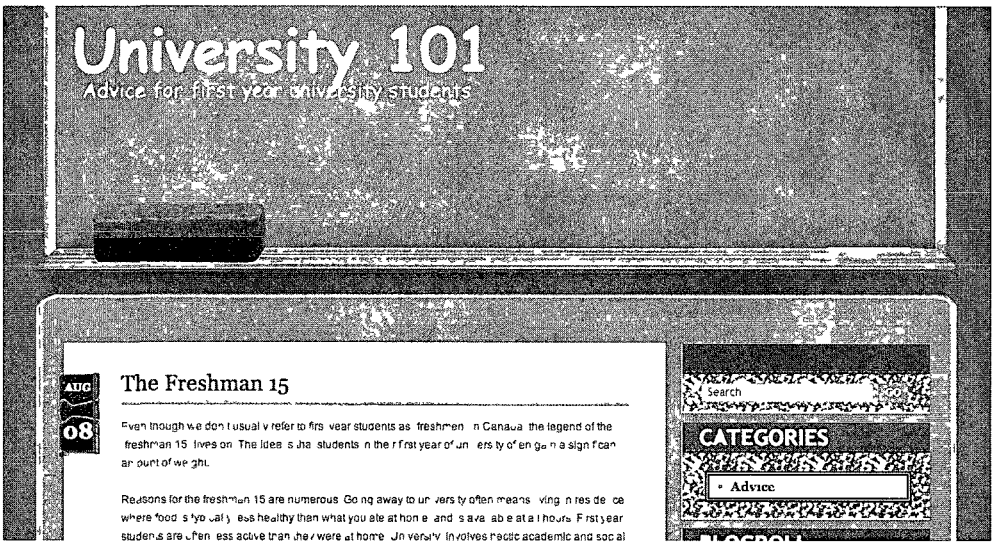


Figure G3 The University 101 blog

Appendix H

Recruitment Notices

For on- and off-campus notices:

Participate in a Website Usability Study and earn $20!

To participate in this study, you must:

- Have regular access to a computer with internet
- Be used to entering usernames and passwords to access websites
- Be able to run specific web software (checked when you contact us)

This study takes place in the lab and online. Well have you complete an initial lab session, then well ask you by email to do short tasks on a couple of websites over the next two weeks before returning to the lab to complete the study and get paid. This study has received clearance by the Carleton University Psychology Research Ethics Board (11-100).

As a participant, youll be paid $20.

Please contact Carleton HCI Study for more details at CarletonHCIStudy@gmail.com

Appendix I

The Mechanical Turk Study

In addition to the lab study, an exploratory replication of the study using a separate participant pool is planned. The participant pool will be obtained from Amazon's Mechanical Turk (MTurk). MTurk is an online marketplace, where workers sign up for voluntary, paid Human Intelligence Tasks or HITs. The idea behind MTurk is that some tasks are faster and easier for a human (or a large group of humans) to perform than a computer. The goal of the replication is to explore the validity of data obtained from MTurk by comparing it to the data obtained through traditional means.

Running the study through MTurk will necessitate several changes to the study procedure. Because MTurk participants are located around the world, the entire study will take place online. Specific changes are outlined below.

Consent and Debriefing The consent and debriefing forms will be presented as webpages. To express their consent to be in the study, the participant will have to type their name and press a button indicating their consent. The debriefing forms will be presented as websites, and the information will also be emailed to participants.

Training The training module in the initial session will be presented via a video that demonstrates how the password systems work. Participants will have the opportunity to watch the video as many times as they feel necessary.

Procedure The procedure followed in the first and third sessions will be delivered as a series of instructions (with links to the appropriate webpages) using the MTurk interface. The emails in the second session will remain unchanged.

Questionnaires All questionnaires in the study will be conducted online using the Limesurvey software, and MTurk participants will be provided with links to the appropriate questionnaires.

Payment A known issue in MTurk is users who attempt to "game" the system by only completing high payment tasks, and completing them with the minimum of effort required.

In order to prevent this kind of participant in our study, the payment for the MTurk study will be $5USD.

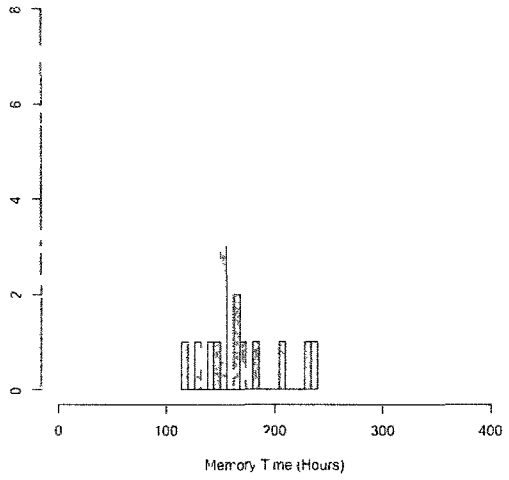
Appendix J

Mechanical Turk Study Results

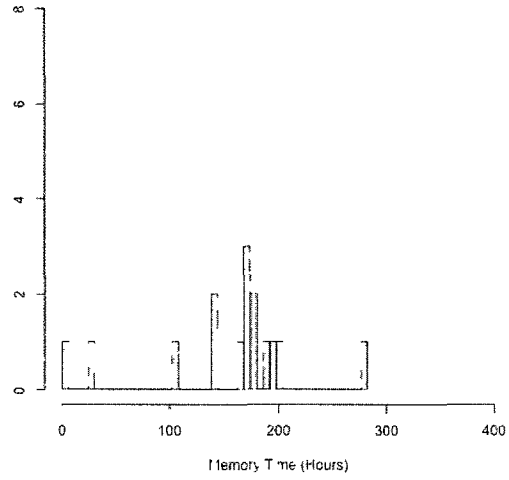
Part of the exploratory analysis was to investigate the use of Amazon Mechanical Turk (MTurk) as a participant pool for usable security experiments. MTurk is an online crowd-sourcing website where workers from around the world can sign up to complete short tasks for payment. While MTurk has been used as a source of participants in several usable security studies (Shay et al., 2010; Komanduri et al., 2011), little work has directly compared results obtaining using MTurk participants to those obtained in more traditional studies. To investigate this, we conducted a replication of our study using Mechanical Turk as a source of participants.

There were 77 participants who completed all parts of the MTurk study: 14 in Blank Passtiles, 15 in Image Passtiles, 14 in Object Passtiles, 18 in Assigned Text, and 16 in Chosen Text. The unequal numbers in each condition are due to inconsistent attrition in the different conditions. The participants in the study ranged in age from 18 to 64, with a median age of 26 years. 30 participants were female and 44 were male. 26 participants from MTurk reported that they were students, from a variety of degree programs and years of study. Most degree programs reported were science or technology programs, but a few participants reported studying in the arts or humanities. Of the remaining participants, the occupations reported varied widely, and included computer professionals, businesspeople, service providers, caring professionals and those not formally employed. 8 participants reported having been in a previous website study, and 5 people reported having previously participated in a password study. 5 participants reported having used a graphical password. Participants came from all over the world, and the largest densities of participants were in the United States (27 participants) and India (26 participants). These participant demographics were very similar to the demographics of the participants in the main study, but there were more students in the main study, and the proportion of genders was approximately reversed.

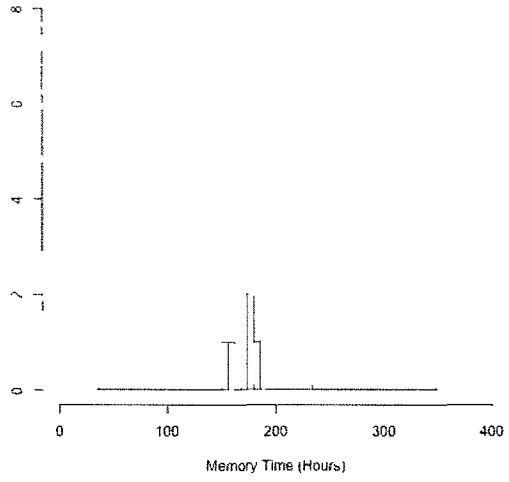
The mean memory times were slightly longer overall in the MTurk study than in our main study, probably because participants were free to complete the final task at any time *after* the



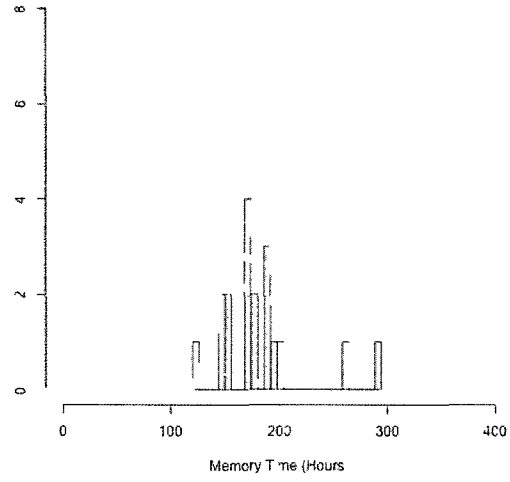
(a) Blank Passtiles



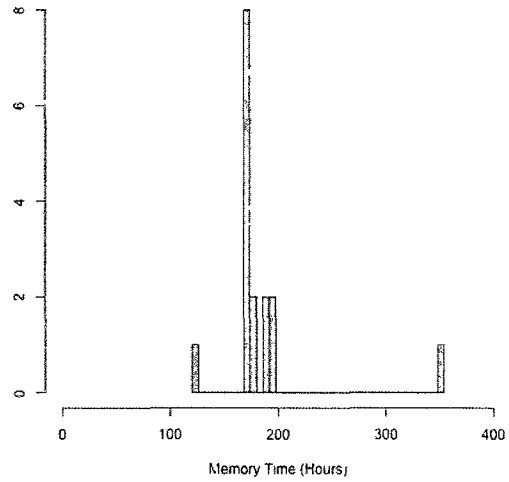
(b) Image Passtiles



(c) Object Passtiles



(d) Assigned Text



(e) Chosen Text

Figure J1 Distributions of memory time for each study condition (MTurk)

Table J1: Descriptive statistics for memory time (in hours). (MTurk)

	Mean	SD	Median	Skewness	Kurtosis
Blank Passtiles	167.38	36.19	157.84	0.88	0.10
Image Passtiles	153.97	67.95	171.28	-0.90	1.80
Object Passtiles	180.63	65.39	175.46	0.72	5.19
Assigned Text	181.76	39.48	172.70	1.50	2.87
Chosen Text	184.37	47.83	171.21	3.11	11.80

final email arrived. As in the main study, Image and Blank Passtiles had the shortest mean memory time, but overall, the mean memory times were similar throughout the conditions (Table J1).

Table J2: t-tests of memory time. (MTurk)

	t	df	p
Assigned Text vs. Blank Passtiles	1.07	29	0.854
Assigned Text vs. Image Passtiles	1.40	22	0.912
Assigned Text vs. Object Passtiles	0.06	20	0.522

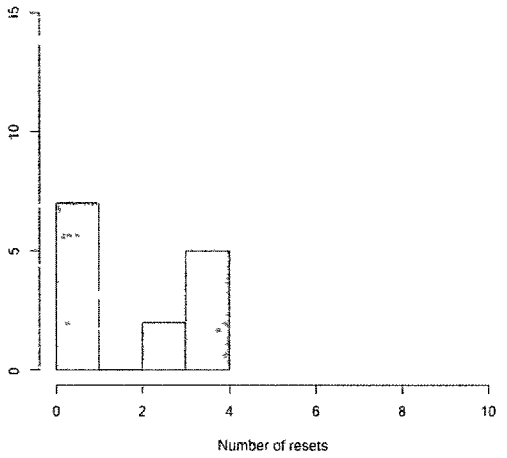
Again similarly to the main study, there were no significant differences in memory time between the assigned text and each of the graphical password conditions (Table J2), or between any of the graphical password conditions (Table J3).

Table J3: ANOVA comparing memory time for the graphical password conditions. (MTurk)

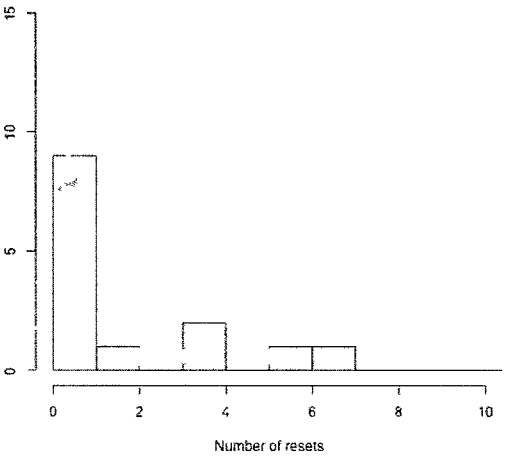
	df	SS	MS	F	p
PassTiles	2	5149.31	2574.65	0.75	0.479
Residuals	40	137257.53	3431.44		

One different finding was in comparing the memory time of chosen text to the memory time of each of the other conditions (Table J4). In the main study, we found that memory time was significantly longer for chosen text than for any of the other conditions, but we did not find this result in the MTurk study. However, in neither case did this appear to represent a large effect.

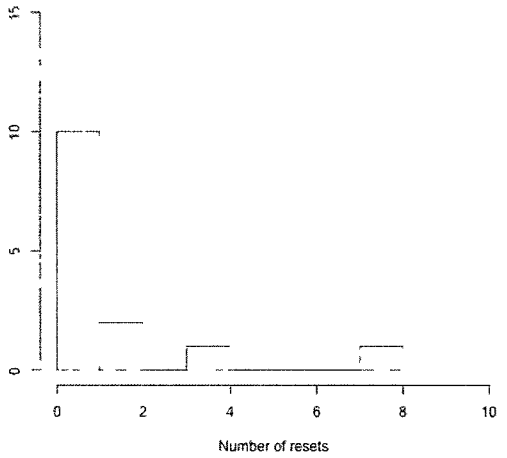
Participants in the MTurk study reset their passwords more often than those in the



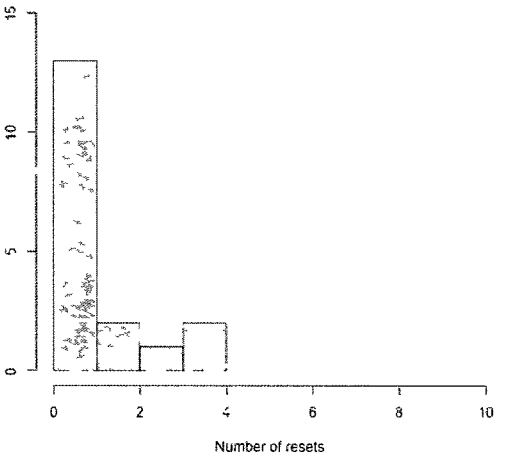
(a) Blank Passtiles



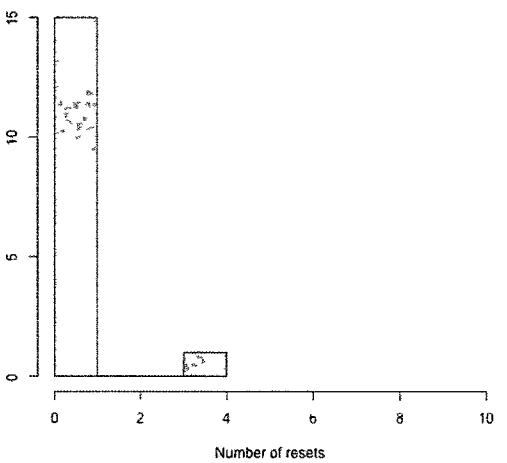
(b) Image Passtiles



(c) Object Passtiles



(d) Assigned Text



(e) Chosen Text

Figure J2. Distributions of resets for each study condition. (MTurk)

Table J4: t-Tests of memory time. (MTurk)

	t	df	p
Chosen Text vs. Blank Passtiles	1.10	27	0.139
Chosen Text vs. Image Passtiles	1.43	25	0.082
Chosen Text vs. Object Passtiles	0.18	24	0.431
Chosen Text vs. Assigned Text	0.17	29	0.432

Table J5: Descriptive statistics for password resets. (MTurk)

	Mean	SD	Median	Skewness	Kurtosis
Blank Passtiles	1.36	1.45	1	0.15	-2.15
Image Passtiles	2.00	3.42	0	2.10	4.63
Object Passtiles	0.86	1.96	0	2.82	8.30
Assigned Text	0.56	1.04	0	1.77	1.88
Chosen Text	0.19	0.75	0	4.00	16.00

in-person study, particularly in the three graphical password conditions (Table J5). There are several possible reasons for this difference. Participants in the MTurk study received less training in the use of the password systems, and may have experienced more trouble in remembering their passwords. Additionally, the MTurk participants may not have attended closely to the instructions, and may not have understood that they would need to remember their passwords for later logins.

Table J6: Wilcoxon tests of password resets comparing the assigned text condition with each of the graphical password conditions. (MTurk)

	U	p
Assigned Text vs. Blank Passtiles	90.50	0.943
Assigned Text vs. Image Passtiles	109.50	0.869
Assigned Text vs. Object Passtiles	124.00	0.548

Compared to the main study, there were more password resets in the Object Passtiles condition, and this affected the differences between conditions. As in the main study, no significance was seen between the Assigned Text condition and either Image Passtiles or Blank Passtiles. However, unlike the main study, no significant difference was seen between Assigned Text and Object Passtiles in the MTurk study (Table J6).

Table J7: Kruskal-Wallis test of resets in the graphical password conditions. (MTurk)

	χ^2	df	p
Kruskal-Wallis chi-squared	1.40	2	0.497

Again similarly to the main study, we did not see any significant differences in password resets between the three graphical password conditions (Table J7). Unlike the main study, there was a significant difference in password resets between chosen text and each of the other conditions (Table J8).

Table J8: Wilcoxon tests of resets comparing the chosen text condition with each of the other conditions. (MTurk)

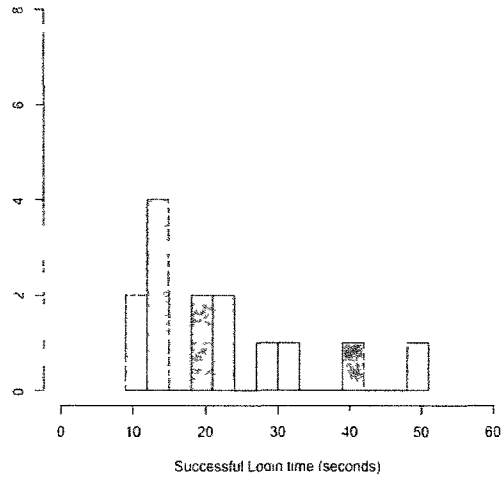
	U	p
Chosen Text vs. Blank Passtiles	64.00	0.005
Chosen Text vs. Image Passtiles	78.50	0.013
Chosen Text vs. Object Passtiles	87.50	0.062
Chosen Text vs. Assigned Text	114.50	0.066

Table J9: Descriptive statistics for login times. (MTurk)

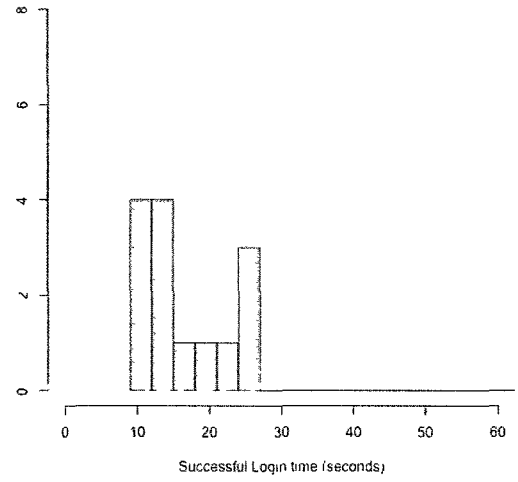
	Mean	SD	Median	Skewness	Kurtosis
Blank Passtiles	21.87	11.41	19.17	1.27	0.93
Image Passtiles	19.95	13.87	14.85	2.80	9.14
Object Passtiles	30.23	9.45	31.08	0.82	1.30
Assigned Text	10.72	5.33	9.54	1.36	2.86
Chosen Text	5.16	1.79	4.71	1.97	5.58

Median login times followed the same pattern in both the main study and the Mturk study, but login times were slightly longer overall in the MTurk study than in the main study (Table J9). This is probably due to participants with slower computers, and the network delays associated with participants in diverse geographic regions.

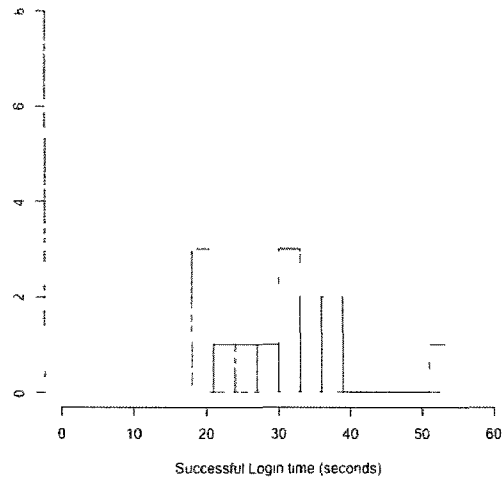
Similarly to the main study, a Kruskal-Wallis test (Table J10) showed significant differences in login time among the five study conditions. Post-hoc pairwise Wilcoxon tests (Table ??) showed significant differences in all pairwise conditions except between Image Passtiles and Blank Passtiles, and between Object Passtiles and Blank Passtiles. This differed from the



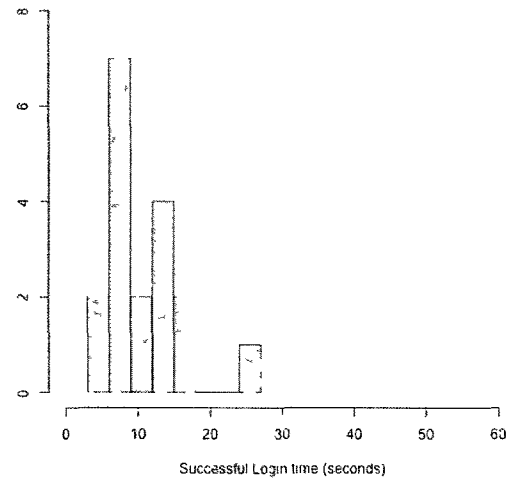
(a) Blank Passtiles



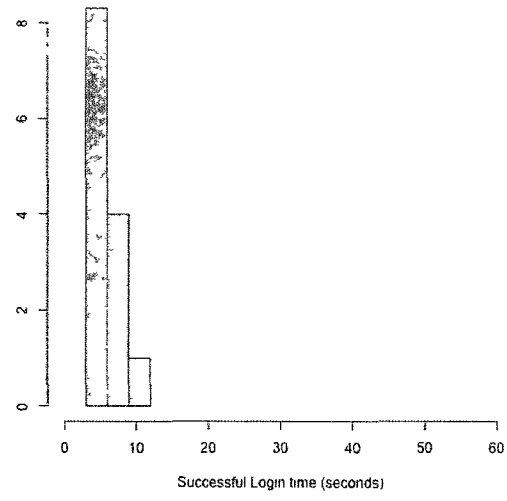
(b) Image Passtiles



(c) Object Passtiles



(d) Assigned Text



(e) Chosen Text

Figure J3. Distributions of login time for each study condition. (MTurk)

Table J10: Kruskal-Wallis test of login times. (MTurk)

	χ^2	df	p
Kruskal-Wallis chi-squared	52.99	4	< 0.001

main study, where all pairwise comparisons were significant except between image passtiles and blank passtiles and between image passtiles and assigned text. This difference arises from the fact that login times for the three graphical password conditions were longer and more similar in the MTurk study than in the main study.

Table J11: Pairwise Wilcoxon tests of login times using Bonferroni adjustment. (MTurk)

	U	p
Blank Passtiles vs. Image Passtiles	122.50	1.000
Blank Passtiles vs. Object Passtiles	50.00	0.274
Blank Passtiles vs. Assigned Text	217.50	0.005
Blank Passtiles vs. Chosen Text	224.00	< 0.001
Image Passtiles vs. Object Passtiles	34.00	0.013
Image Passtiles vs. Assigned Text	214.50	0.043
Image Passtiles vs. Chosen Text	238.00	< 0.001
Object Passtiles vs. Assigned Text	247.00	< 0.001
Object Passtiles vs. Chosen Text	224.00	< 0.001
Assigned Text vs. Chosen Text	257.50	0.001

As in the main study, participants were asked in the post-test questionnaire whether they had written down any of their passwords. Figure J4 shows the participants in each condition in the turk study who reported writing down their passwords. More participants reported writing down their passwords in the MTurk study than in the main study. In the MTurk study, participants were not instructed not to write their passwords down, partly for ecological validity, and partly to avoid false replies on the post-test questionnaire. Unlike in the main study, a Chi-squared test showed that there was no significant difference ($\chi^2(3) = 6.72, p = 0.081$) in the number of password recordings between the four study conditions. This difference was likely due to the higher proportion of participants in the MTurk study in the Object Passtiles condition who wrote their passwords down.

The results obtained in the MTurk study look tentatively comparable to those obtained in

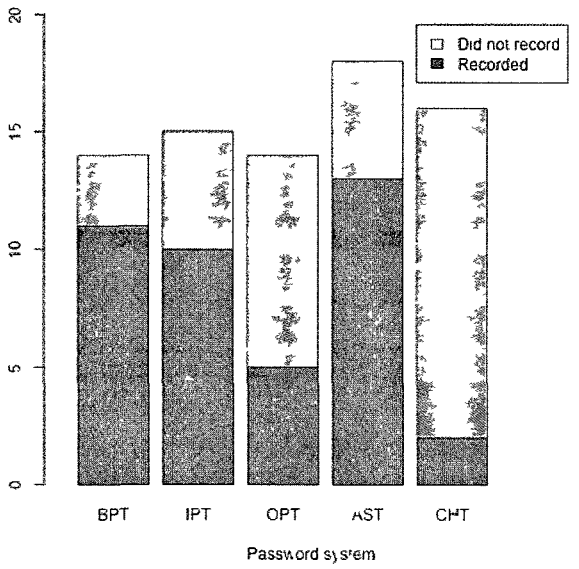


Figure J4 Frequency of reported password recording. (MTurk)

the main study, but there are a few differences in the way the studies were conducted that may be affecting the results. The MTurk participants were not able to receive the in-person training on the password system. While the in-person participants were offered a brief explanation of the password system, and were allowed to become familiar with the system through a simple training page, MTurk participants were merely told (via email) that the password system might be unusual, and a link to an explanatory help page was easily available. MTurk participants also did not have the opportunity to ask questions about the password system in real time, though there was a help email address that some participants used to ask questions.

Another factor affecting the results of the two studies was that the experiments did not take place simultaneously. The MTurk study began as the in-person study was concluding, and the bulk of it took place after the main study had finished. Also contributing to the differences between the in-person study and the mTurk study were the social pressures present in the main study. MTurk participants had the anonymity of the internet, and this affected the extent to which they felt pressured to remain in the study, and perform well. In contrast, participants in the main study were known to the experimenters, and were probably more influenced by

social rules about reciprocity and politeness. As well, participants in the main study may have been more eager to perform well to please the experimenters.

One phenomenon we noticed in running the MTurk study was that many more people began the study than finished it. In the MTurk replication of the main study, 119 people completed Session 1, but only 77 people completed Session 3. In all conditions, there was a steep drop (approx 40% in the assigned password conditions, 23% in the chosen text condition) in the number of people who completed Session 1 and the number of people who completed Session 2, but there was little to no drop between Session 2 and Session 3. It is difficult to fully understand the reasons for this attrition, but we speculate that since Session 1 was more difficult, participants assumed that the remaining parts of the study would be similarly difficult. In fact, there was less work associated with Sessions 2 and 3, and the payment was comparatively higher for Session 2. We also note that fewer people dropped out of the chosen text condition, which was more straightforward for users.

In addition to the attrition between the study sessions, there was also self-selection in the MTurk study. There was a large difference between the number of people who received the instructions for Session 1 (indicating their interest in participating) and the number of people who finished Session 1. In Blank and Image Passtiles, only about half (51%) of the people who received instructions for Session 1 completed the session. The percentages increased for Object Passtiles (61%), Assigned Text (72%), and Chosen Text (84%). However, although attrition was an issue for the MTurk study, these differences were not significant.

The patterns of attrition and self-selection seen in the MTurk data suggest that participants who find the study difficult are inclined not to complete the study. This may mean that the results obtained using Amazon Mechanical Turk are somewhat better compared to what we might see in an in-person study. However, this is not to say that there are not elements of self-selection, attrition and bias in traditional studies.

Overall, our experiences in running the MTurk study suggest that research using this resource is possible, but that researchers should be aware of the possible differences in results obtained through MTurk, and should be sure to examine their study designs for possible issues and confounds before broadly interpreting the results.