Memory Retrieval and Graphical Passwords

Elizabeth Stobert
School of Computer Science
Carleton University
Ottawa, Canada
elizabeth_stobert@carleton.ca

Robert Biddle
School of Computer Science
Carleton University
Ottawa, Canada
robert_biddle@carleton.ca

ABSTRACT

Graphical passwords are an alternative form of authentication that use images for login, and leverage the picture superiority effect for good usability and memorability. Categories of graphical passwords have been distinguished on the basis of different kinds of memory retrieval (recall, cued-recall, and recognition). Psychological research suggests that leveraging recognition memory should be best, but this remains an open question in the password literature. This paper examines how different kinds of memory retrieval affect the memorability and usability of random assigned graphical passwords. A series of five studies of graphical and text passwords showed that participants were able to better remember recognition-based graphical passwords, but their usability was limited by slow login times. A graphical password scheme that leveraged recognition and recall memory was most successful at combining memorability and usability.

Categories and Subject Descriptors

K.6.5 [**Management of computing and information systems**]: Security and protection—*authentication*

General Terms

Human Factors

Keywords

Usable security, graphical passwords, authentication, human memory

1. INTRODUCTION

Text passwords are a widely used form of authentication, but users often select easily guessable passwords [10]. Graphical passwords are a proposed alternative to text passwords that have been shown to have good usability and security properties [2]. They are now beginning to be widely deployed (e.g., Android Pattern Unlock, Windows 8 Picture Passwords). They can be classified into three categories, distinguished by the type of memory retrieval leveraged in the password scheme. These categories are: recall-based, cued recall-based, and recognition-based graphical passwords. To date, there is little evidence as to which approach is best. This paper addresses this open question by investigating how different kinds of memory retrieval affect the usability and memorability of random assigned graphical passwords.

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 proposed systems ask users to draw a password image [14], click different places on a picture [29], or identify pictures of faces [23]. Graphical passwords leverage the picture superiority effect [22], which says that humans remember images better than they remember textual information.

The different categories of graphical passwords 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 they must click correct points on the image to log in. Recognition-based graphical passwords present users with an array of images, and the user must choose the correct images to log in. Although psychological research has shown that recognition memory is superior to recall [12], the question of whether recognition-based graphical passwords are more memorable than other types of graphical password remains unresolved. In this paper, we compare the three categories of graphical password in a study based in real world usage.

Existing graphical password systems are difficult to compare, since they use different types of input and have differing levels of security. We designed a new graphical password system to allow easy comparison of the different types of memory retrieval, and conducted a set of five studies with a total of 336 participants to investigate how leveraging different kinds of memory retrieval can affect the usability and memorability of assigned graphical passwords. We found that users were better able to remember recognition-based passwords, but the associated login times were too slow for real-life use. In our final study, we modified the password system to allow users to take advantage of both recognition and recall memory, and these passwords were faster to enter and memorable for users. We propose that this combination may be the best approach for graphical passwords.

The rest of the paper proceeds as follows: first we provide background on the psychological research about the

Symposium on Usable Privacy and Security (SOUPS) 2013, July 24–26, 2013, Newcastle, UK
.

picture superiority effect and memory retrieval. We introduce *PassTiles*, a new graphical password system designed to allow direct comparison of different types of memory retrieval. We then describe our study methodology and the results of our studies.

2. BACKGROUND

Graphical passwords are said to leverage the *picture superiority effect* [22], or the finding that people have better memory for images than words. The picture superiority effect is seen in tests of both recall and recognition.

Paivio's *dual coding theory* [21] postulates that the brain has separate mechanisms for remembering image-based 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 [17] as the *distinctiveness account.* Nelson, Reed, and McEvoy [19] 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 supported by evidence that visual similarity in images leads to decreases in the picture superiority effect [20].

The levels-of-processing approach [5] 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 [18] found evidence that the picture superiority effect is related to the different processing applied to images and words.

2.1 Memory Retrieval

Different graphical password schemes leverage different types of memory retrieval through their design. The differing kinds of retrieval affect not only memorability, 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 [13]. 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. Recognition is almost always found to be superior to recall [12], and there are several theories that attempt to explain the differences.

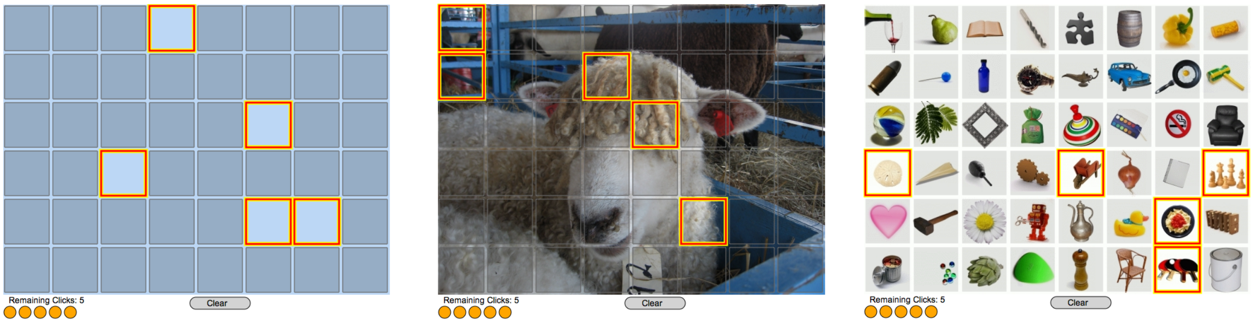
One of the most prominent theories of retrieval is the *generate-recognize theory* [1]. This theory posits that retrieval is a two-step process, consisting of both generation and recognition phases. For example, there are two phases to retrieve the memory of a word. In the generate phase, long term memory is searched, and a list of candidate words is formed. Then in the recognize phase, the words in the list are evaluated to see if they can be recognized as the sought-out word. 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.

The generate-recognize theory explains some of the differences between recognition and recall memory [28]. 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 [28]. Most notably, the theory has difficulty explaining the success and failure of some kinds of cueing. Studies have shown that it is best when the cue for a studied list item is consistent, because where it is changed (e.g., "sail" vs. "gravy" for boat) subjects have a harder time recalling the appropriate word. However, in studies of recognition, even non-studied cues can aid the retrieval of unrecognized words.

In reaction to research about the effects of unlearned cues on recall memory, Tulving and Thompson [25] posited the *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 the 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.

One of the major differences between the generate-recognize theory and the encoding specificity theory is the assumed level of complexity of the retrieval process. Generate-recognize 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, the 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.

A further distinction made in the memory literature is between *free recall* and *cued-recall.* 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* [9] 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



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

Figure 1: Password creation interfaces for the three graphical password schemes used in the study.

the two items occur together. In contrast, encoding specificity theory [25] says that the most effective cues are the cues that are present at the time of remembering.

2.2 Types of Graphical Passwords

There are many types of graphical passwords [2], but we focus on the categorization often made by distinguishing the kind of memory leveraged by the scheme [7]:

Recall-based: Also known as drawmetric, these systems ask the user to reproduce a drawing on a grid. Example schemes include Draw-a-Secret (DAS) [14].

Cued recall-based: In these passwords, the user is asked to accurately click on points on an image (also referred to as locimetric or click-based graphical passwords). Example schemes include PassPoints [29].

Recognition-based: Also known as cognometric, these schemes ask the user to recognize and identify images belonging to their set of password images from a set of distractor images. An example scheme is PassFaces [23].

Most drawmetric graphical password systems leverage the picture superiority effect by using a grid-drawing exercise, but 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*. Different locimetric graphical password systems utilize different methods of password entry, but all locimetric graphical password systems provide users with visual cues to help them more easily recall and distinguish their passwords, thus leveraging *cued-recall*. 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. They leverage *recognition* memory by explicitly displaying all possible choices to the user, and expecting them to recognize the correct option.

Most graphical password systems allow users to select their own passwords. However, work on user-choice in graphical passwords [6, 4, 26, 27] has shown that users tend to choose predictable passwords that can be exploited in a dictionary attack. Assigning random passwords protects against this kind of attack. In some cases, the predictability of user-chosen passwords can significantly affect the security of the system. Davis, Monrose, and Reiter [6] showed that users tended to choose Passfaces passwords with strong gender and attractiveness biases, and these biases opened the passwords to attack. Following these results, Passfaces changed to assigned random passwords.

Since it is not clear that users choose equivalently predictable passwords in the three types of graphical password, a fair comparison of the categories would need to assign random passwords. This would ensure that the security of the systems was equivalent, allowing a valid comparison of memorability and usability. Assigning passwords presents different challenges for the different types of graphical password. The complexity and granularity of existing drawmetric password systems makes assigning passwords difficult. Clearly conveying the appropriate details is problematic in systems where there are many subtle nuances to password entry that can be difficult to convey to users. Existing locimetric schemes such as PassPoints [29] require precise input that can create 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 passwords are easily assigned by showing the user their set of password images at password creation. In the next section, we present our approach that allows all three categories of passwords to be assigned in a similar way.

3. PASSTILES

PassTiles (Figure 1) is a new graphical password system that we created for use in this research. We designed PassTiles to be able to compare different types of memory retrieval within the same password system. To be able to easily assign passwords to users, we combined features of DAS, PassPoints and Passfaces. In PassTiles, the user is presented with a grid of password tiles and their password consists of five password tiles, which are randomly assigned by the system. To log in, the user must click on the correct password tiles. The order of tile entry is not significant.

Blank PassTiles (Figure 1(a)) is a version of PassTiles that uses a grid with a blank background. Having a blank background makes password retrieval a free recall task for the user, similar to DAS. The patterns formed by the password tiles are similar to a mosaic and thus leverage visual memory similarly. Image PassTiles (Figure 1(b)) superimposes the grid of password tiles over an image, to allow users to remember their password in relation to the background image. Similar to PassPoints, Image PassTiles takes advantage of cued-recall memory, and the background image provides users with a cue to help them remember their password tiles. Object PassTiles (Figure 1(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 ob-

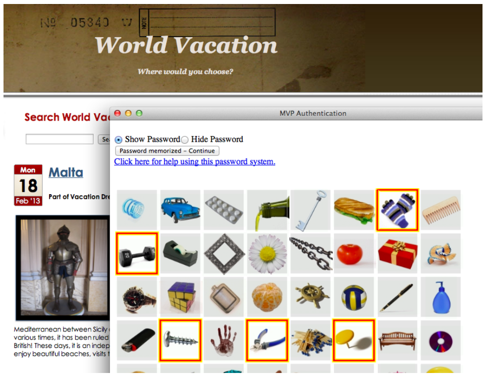

№ 05340 W
NOTE
World Vacation
Where would you choose?
Search World Va
Se
Mon
18
Feb '13
Malta
Part of Vacation Dr
Mediterranean between Sicily
various times, it has been ruled
British! These days, it is an indep
enjoy beautiful beaches, visits t
MVP Authentication
● Show Password ○ Hide Password
Password memorized – Continue
Click here for help using this password system.

Figure 2: One of the three websites used in the studies, showing the PassTiles password creation screen.

jects that the user must click on to login. 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. Without shuffling, users would be able to rely on recall to remember the positions of the password tiles. Our usage model was a standard desktop or laptop computer, with a colour display and a mouse or trackpad for input.

In the context of graphical passwords, we feel it is important that cued-recall involve a cue that is specific to the password being entered, rather than a cue that would apply to any (and all) passwords. This distinguishes the type of recall leveraged in Blank PassTiles and Image PassTiles. To make certain all of the cues (or objects to be recognized) were distinct, the image sets in our studies never overlapped.

Our 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 was easy to learn. Learnability was emphasized because we wished the schemes to be viable in a real-world situation, and not require participants to undergo extensive training. All PassTiles schemes were designed to help the user understand the scheme and practice entering their password as part of the password creation process.

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.

4. STUDY DESIGN

The goal of our studies was to explore 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? To investigate this question, we conducted a series of five studies. Each study was a between-subjects study, where the password scheme varied by condition and participants were randomly assigned to one condition. Participants created and used passwords on different websites over the course of one week. Although the study duration was limited, we felt it was appropriate to explore differences in this time period before more extensive testing was justified. Our studies were approved by the Carleton University Research Ethics Board.

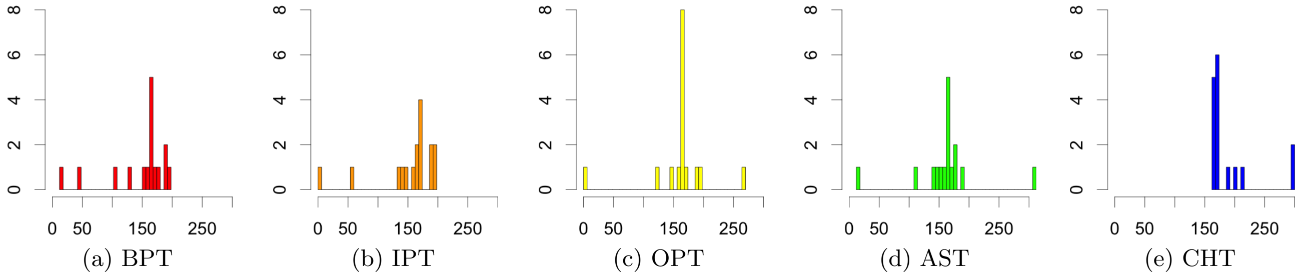
Table 1: Password spaces for the configurations of the password systems used in the study.

Password	Configuration	Password
system		Space
BPT	8 × 6 grid, length 5	$log_2\binom{6\times 8}{5} = 21$ bits
IPT	8 × 6 grid, length 5	$log_2\binom{6\times 8}{5} = 21$ bits
OPT	8 × 6 grid, length 5	$log_2\binom{6\times 8}{5} = 21$ bits
AST	36 characters, length 4	$log_2 36^4 = 21$ bits
CHT	36 characters, length 4	$log_2 36^4 = 21$ bits

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 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. Of course, novel password systems may signal the participant that the passwords are being studied, but this is unavoidable. The MVP framework [3] was used to implement the password systems on real websites and collect detailed usage data. 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 implement graphical password systems. Figure 2 shows one of the websites used in the study, which were created, hosted, and maintained by us.

The five study conditions were: Blank PassTiles (BPT), Image PassTiles (IPT), Object PassTiles (OPT), Assigned Text (AST), and Chosen Text (CHT) (we use the abbreviations in tables and figures). Three conditions used PassTiles, and the remaining two study conditions used text passwords to provide 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 current practices and the level of security that users choose for themselves. For all of the password systems used in the study, account creation included two steps: password creation, where the user was assigned or, in the case of Chosen Text, 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 [11] 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 studies. Table 1 shows the the configuration and theoretical password spaces for the password schemes used in the studies. For example, in the PassTiles conditions, $log_2\binom{6\times 8}{5} = 21$ bits. In the case of Chosen Text, the passwords were user-chosen but the table shows the theoretical password space.



(a) BPT (b) IPT (c) OPT (d) AST (e) CHT

Figure 3: Distributions of memory time (in hours) by number of participants for each study condition (Study 1).

Each study took place in three sessions:
Session 1: Participants were briefly trained on the password system and were introduced to each of the three websites in the study, creating and confirming a password for each site. They completed a demographics questionnaire before completing a short task (such as commenting on an article) on each site. These tasks necessitated logging in to each website.
Session 2: Participants were 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, and the sixth day. 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.
Session 3: Session 3 took place one week after session 1. Participants completed a final task on each website, and filled out the post-test questionnaire, which asked them about their experiences using the password system.

The five studies conducted followed the basic procedure outlined here. In Study 1, participants were recruited from our local community, and sessions 1 and 3 were conducted in-person. Studies 2 to 5 were conducted online, using Amazon's Mechanical Turk as a participant pool. Study 4 used a modified procedure, where session 2 was omitted and session 3 was conducted after only 2 days. We initially used in-lab studies to tightly control the method and observe users, and later used online studies to gain access to larger numbers and more diverse participants. The five studies followed up on each other, but we never made direct statistical comparisons between studies.

The independent variable in the studies was the password system used. The effect should indicate the usability of the scheme, including both memorability and other relevant factors. The dependent variables that were used as a measure of memorability were the average length of time that a password was remembered, the number of password resets, and the time to login. As an indicator of password memorability, we measured the *memory time*, or the average length of time that a participant remembered their password. For each account, memory time was measured as the greatest length of time between a password creation and the a successful password login (using the same password). This is an inherently conservative measure, reflecting our emphasis on ecological validity, since participants may have remembered their passwords longer than we were able to measure. It can also be influenced by when participants chose to return to tasks we set, but we have no reason to believe this would affect conditions differently. The second measure of memorability was the number of password resets per account. The MVP system allows users to reset their passwords without experimenter intervention when forgotten. Finally, we measured login time as an indicator of password system usability. Login time was measured for successful password attempts from the time that the entry window appeared on screen until the website verified the entry attempt as successful. The number of login attempts was unlimited.

5. RESULTS

5.1 Study 1

Study 1 was conducted in a hybrid format, where session 1 was conducted in person, session 2 took place online, and session 3 was in person. The goal of the study was to investigate how the different types of memory retrieval affected the usability and memorability of assigned graphical passwords.

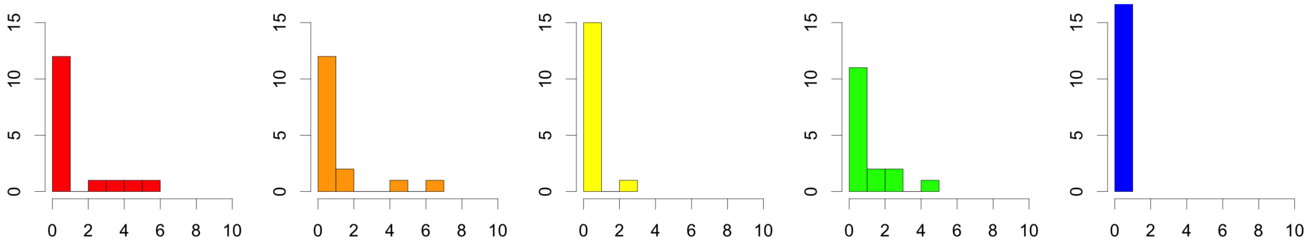
Study 1 had 81 participants (45 female), recruited from both the university and the local community. Participants ranged in age from 18 to 62, with a median age of 24. 53 participants were students from a broad range of degree programs and levels, and the remaining participants were occupied in a broad variety of fields. None of the students were studying or working in the field of computer security.

We had three hypotheses, each applying to one of the dependent variables studied:

H1(a): Memory time would be significantly shorter for Assigned Text than for the three graphical password conditions.
H1(b): There would be significant differences in memory time among the three graphical password conditions.
H1(c): Memory time would be significantly longer for Chosen Text than for any of the assigned password conditions.

H2(a): There would be significantly more password resets in Assigned Text than in any of the three graphical password conditions.
H2(b): There would be significant differences in the number of password resets among three graphical password conditions.
H2(c): There would be significantly fewer password resets in Chosen Text than in any of the assigned password conditions.

H3: There would be significant differences in login times among the five study conditions.



(a) BPT (b) IPT (c) OPT (d) AST (e) CHT

Figure 4: Distributions of resets by number of participants for each study condition (Study 1).

5.1.1 Hypothesis 1: Memory time

Table 2: Memory time statistics (hours) (Study 1).

	Mean	SD	Median	Skewness	Kurtosis
BPT	147.00	50.57	164.14	-1.77	2.57
IPT	150.76	51.90	167.63	-2.16	4.61
OPT	160.71	51.96	166.90	-1.63	7.05
AST	160.67	56.80	166.57	-0.02	5.60
CHT	190.62	44.08	168.73	2.09	3.35

Table 2 shows descriptive statistics for the memory time variable. To aggregate across the three websites used in the study, we took the mean of each participant's memory time for each website. Median memory time ranged between 164.14 hours in Blank PassTiles and 168.73 hours in Chosen Text. The total duration of the study was 7 days (approximately 168 hours), so this result shows that most participants were able to remember their passwords for the entire study. We excluded one participant from the analysis because they had to leave town during the 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. We were surprised to find that all participants had been able to remember their passwords so long.

Histograms of the distributions of memory time (in hours) (Figure 3) for each of the five conditions suggested that the distributions are approximately normal. Some conditions were leptokurtotic, but the measures of skewness and kurtosis (Table 2) indicated that this was unlikely to affect the results and we conducted further analysis using ANOVA and t-tests. We follow this procedure in all tests we report in this paper, using parametric tests (ANOVA and t-tests) where appropriate, and otherwise apply ordinal tests (Kruskal-Wallis and Wilcoxon (Mann-Whitney) tests).

H1(a): We conducted three one-sided t-tests, each comparing a graphical password condition to Assigned Text, and found no significant differences in memory time between Assigned Text and any of the graphical password conditions. This provided no evidence that participants were able to remember graphical passwords longer than assigned text passwords. Since these comparisons were hypothesized *a priori*, we did not correct for multiple tests.

H1(b): A one-way ANOVA of memory time showed no significant differences in memory time among the three graphical password conditions.

H1(c): We conducted a set of four *a priori* t-tests comparing memory time for Chosen Text with 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 ($t(29) = 2.34, p = 0.013$), and between Chosen Text and Object PassTiles ($t(29) = 1.76, p = 0.045$). We suspect that this result reflects a chance occurrence where more Chosen Text participants returned after 7 days, artificially creating a difference in memory times.

5.1.2 Hypothesis 2: Resets

Table 3: Password reset statistics (Study 1).

	Mean	SD	Median	Skewness	Kurtosis
BPT	0.88	1.67	0	1.70	1.61
IPT	0.75	1.73	0	2.55	5.98
OPT	0.12	0.50	0	4.00	16.00
AST	0.62	1.15	0	2.07	4.26
CHT	0.00	0.00	0	0.00	-3.66

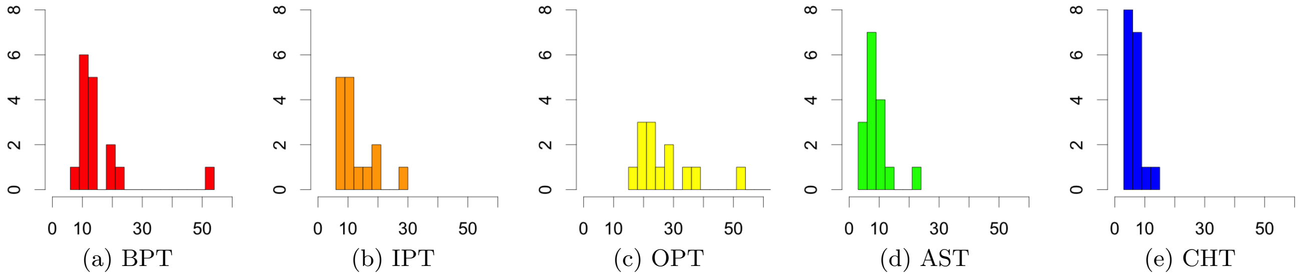
As a measure of the memorability of the password systems, we recorded the number of password resets. For each participant, we took the total number of resets on all websites. Participants were free to reset their passwords at any time during the at-home sessions of the study.

The median number of resets was 0 in all conditions, indicating that most participants never reset any of their passwords. Table 3 shows descriptive statistics for password resets. Figure 4 shows the distributions of password resets for each condition. As seen in the histograms, the distributions of resets were skewed and kurtotic, making parametric tests unsuitable. Wilcoxon tests were conducted in place of t-tests, and Kruskal-Wallis tests were used in place of one-way ANOVAs.

H2(a): Wilcoxon tests showed no significant difference in the number of password resets between Assigned Text and either Image PassTiles or Blank PassTiles. However, participants reset their passwords significantly more often in Assigned Text than in Object PassTiles ($U = 159.50, p = 0.043$). Again, since these comparisons were hypothesized *a priori*, we did not correct for multiple tests.

H2(b): A Kruskal-Wallis test showed no significant difference in the number of password resets among the three graphical password conditions, providing no evidence that participants reset their passwords differently in any of the PassTiles conditions.

H2(c): We conducted four *a priori* one-way Wilcoxon tests, and found a significant difference in the number of



(a) BPT (b) IPT (c) OPT (d) AST (e) CHT

Figure 5: Distributions of login time (in seconds) by number of participants for each study condition (Study 1).

password resets between Chosen Text and each of Blank PassTiles ($U = 102.00, p = 0.017$), Image PassTiles ($U = 102.00, p = 0.017$), and Assigned Text ($U = 93.50, p = 0.008$). This indicated that in all but the Object PassTiles condition, participants reset their passwords more often in the assigned text condition.

5.1.3 Hypothesis 3: Login times

Table 4: Login time statistics (seconds) (Study 1).

	Mean	SD	Median	Skewness	Kurtosis
BPT	15.74	10.48	13.38	3.00	10.19
IPT	12.65	6.14	9.33	1.38	1.52
OPT	34.61	21.79	24.25	1.74	2.03
AST	9.06	3.92	8.28	2.12	5.76
CHT	6.35	2.41	6.00	1.71	3.87

Table 4 shows descriptive statistics for login times. Mean login times varied widely across conditions, ranging from 6.35 seconds in Chosen Text to 34.61 seconds in Object PassTiles. 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.

As seen in Figure 5, the distributions of login times were right skewed and leptokurtotic, making the use of parametric tests inappropriate.

Table 5: Pairwise Wilcoxon tests of login times using Bonferroni adjustment (Study 1).

	U	p
BPT vs. IPT	153.00	1.000
BPT vs. OPT	22.00	0.001
BPT vs. AST	213.00	0.014
BPT vs. CHT	260.00	< 0.001
IPT vs. OPT	13.00	< 0.001
IPT vs. AST	169.50	0.527
IPT vs. CHT	230.50	0.001
OPT vs. AST	236.00	< 0.001
OPT vs. CHT	255.00	< 0.001
AST vs. CHT	209.50	0.085

A Kruskal-Wallis test showed significant differences in login times ($\chi^2(4) = 51.31, p < 0.001$) between the different study conditions. Post-hoc pairwise Wilcoxon tests using a Bonferroni adjustment (Table 5) showed that it took participants significantly less time to log in using text passwords, and significantly longer to log in using Object PassTiles.

5.1.4 Hypothesis Testing Summary

The results of Study 1 showed that login times were significantly longer in the recognition condition (Object PassTiles). However, we found no significant differences in memorability between the different password systems. We were surprised by the lack of differences, and address this issue below.

5.1.5 Security Analysis

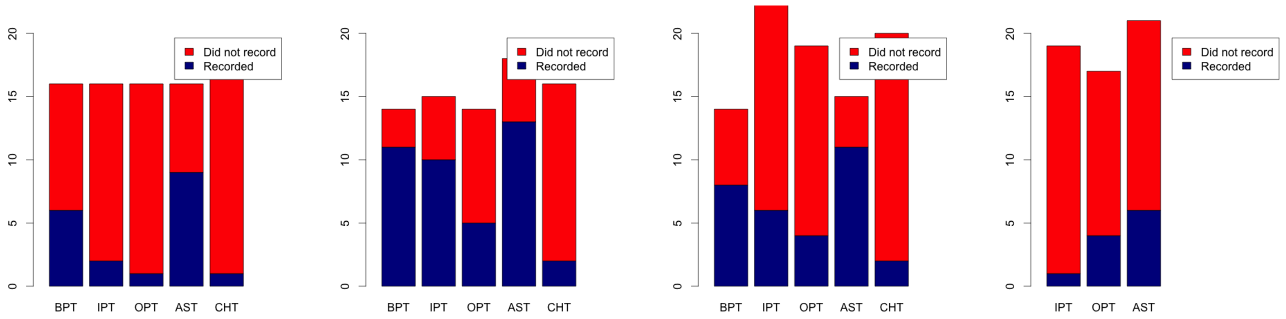
In all conditions but Chosen Text, passwords were randomly assigned, making the effective password space equal to the theoretical password space and pre-determining the security of the password system against guessing attacks. However, the password security in Chosen Text was largely determined by participants'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

When creating their passwords, participants were limited to 4 characters but were allowed to use any character set. In order to examine the security of the passwords chosen by participants in this condition, we looked at the incidence of password reuse, and at the use of patterns (e.g., dictionary words) in password selection.

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 Chosen Text, 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.

The presence of dictionary words and common substitutions further narrows the space. Using the free dictionaries from the *John the Ripper* password cracking program [8], we were able to guess 8 passwords 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 therefore 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 Chosen Text, 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



(a) Study 1 (b) Study 2 (c) Study 3 (d) Study 4

Figure 6: Stacked bar plots showing frequency of reported password recording.

differing security levels, we feel that it is inappropriate to compare the memorability and usability of Chosen Text to the assigned password conditions. In the remainder of the paper, we report statistics for Chosen Text, but do not consider any comparison of memorability with other conditions to be fair.

5.1.6 Password Recording

We were surprised by the results we found for memorability. We had expected to see differences between conditions, and had not expected that most participants would be able to remember their passwords for the duration of the study. 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, we had no means of controlling their behaviour once they left the lab. The post-test questionnaire asked participants 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.

The number of participants who reported recording their passwords in each condition ranged from nine people in Assigned Text to just one person in Chosen Text and Object PassTiles. Figure 6(a) shows a stacked barplot of the counts of password recording (and non-recording) in each condition. The corresponding percentages are: Blank PassTiles= 38%, Image PassTiles= 13%, Object PassTiles= 6%, Assigned Text= 56%, and Chosen Text= 6%.

We used a chi-squared test 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, and post-hoc pairwise chi-squared tests using a Bonferroni adjustment showed that the only significant difference in instances of password recording was between Assigned Text and Object PassTiles ($\chi^2(1) = 7.13, p = 0.046$).

The proportion of participants who reported that they had written down their passwords in the free 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. This result sheds doubt on the hypothesis testing presented above. However, the differences in reported password recording across conditions may suggest that remembering recall-based passwords was more difficult for participants.

Our exploration of password recording in Study 1 was done post-hoc, without any hypothesis. However, the results of the analysis convinced us that this issue was worth exploring more explicitly. In later studies, we considered the issue *a priori*, and hypothesized that stronger password cues would lead to fewer reported password recordings.

5.2 Study 2

The goal of Study 2 was to replicate Study 1 with a larger sample size and more diverse participants. We conducted Study 2 using Crowdflower[1] as an outsourcing service for Amazon's Mechanical Turk[2] (MTurk). MTurk has been posited as an easily available source of participants for usable security experiments [16, 15], but little work has investigated the comparability of results obtained on MTurk to those obtained through more traditional means. We replicated almost all the details of our in-person study in our MTurk study, making the results directly comparable. Only a few changes to the procedure were necessary. MTurk participants received training on the websites through an instructional webpage, instead of an in-person explanation. Also, instead of paying participants for the entire study at completion, we paid participants for each session as it was completed.

The results of Study 2 were very similar to those of Study 1, and for reasons of brevity, we do not include results of Study 2 (and subsequent studies) at the same level of detail as Study 1.

Study 2 had 77 participants and the demographics 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).

Memory times were slightly longer overall in Study 2 (Table 6), probably because participants were free to complete the final task at any time *after* the final email arrived. However, the distribution of memory time was similar to Study 1. As in Study 1, memory times were similar throughout the conditions, and there were no significant differences in memory time between Assigned Text and each of the graphical password conditions, or among the three graphical password conditions.

[1] `http://www.crowdflower.com`

[2] `http://www.mturk.com`

Table 6: Descriptive statistics for memory time (in hours)(Study 2).

	Mean	SD	Median	Skewness	Kurtosis
BPT	167.38	36.19	157.84	0.88	0.10
IPT	153.97	67.95	171.28	-0.90	1.80
OPT	180.63	65.39	175.46	0.72	5.19
AST	181.76	39.48	172.70	1.50	2.87
CHT	184.37	47.83	171.21	3.11	11.80

Table 7: Descriptive statistics for password resets (Study 2).

	Mean	SD	Median	Skewness	Kurtosis
BPT	1.36	1.45	1	0.15	-2.15
IPT	2.00	3.42	0	2.10	4.63
OPT	0.86	1.96	0	2.82	8.30
AST	0.56	1.04	0	1.77	1.88
CHT	0.19	0.75	0	4.00	16.00

Participants in Study 2 reset their passwords more often than those in Study 1, particularly in the three graphical password conditions (Table 7). This was possibly because they received less mandatory training in the use of the password systems, and may have experienced more trouble in remembering their passwords. Study 2 had more password resets in Object PassTiles, and this affected the differences between conditions. In Study 2, there were no significant differences between Assigned Text and any of the graphical password conditions. As in Study 1, we saw no significant differences in password resets between the three graphical password conditions.

Table 8: Login time statistics (seconds) (Study 2).

	Mean	SD	Median	Skewness	Kurtosis
BPT	21.87	11.41	19.17	1.27	0.93
IPT	19.95	13.87	14.85	2.80	9.14
OPT	30.23	9.45	31.08	0.82	1.30
AST	10.72	5.33	9.54	1.36	2.86
CHT	5.16	1.79	4.71	1.97	5.58

Median login times followed the same pattern, but were slightly longer overall in Study 2 than in Study 1 (Table 8). This is probably due to participants with slower computers, and the network delays associated with participants in diverse geographic regions. As in Study 1, a Kruskal-Wallis test showed significant differences in login time among the five study conditions ($\chi^2(3) = 28.95, p < 0.001$). Post-hoc pairwise Wilcoxon tests showed significant differences in all pairwise comparisons except between Image PassTiles and Blank PassTiles, and between Object PassTiles and Blank PassTiles.

Figure 6(b) shows the participants in each condition in Study 2 who reported writing down their passwords. More participants reported recording their passwords in Study 2 than in Study 1. In Study 2, participants were not instructed not to write their passwords down, partly for ecological validity, and partly to avoid false responses on the post-test questionnaire. Again, a chi-squared test showed a significant difference in the number of password recordings ($\chi^2(4) = 19.69, p < 0.001$) between the four assigned password conditions. In our added hypothesis, we speculated that we might again see the relationship between the strength of cues and password recording. In Study 2, the ordering of password write-down frequency was the same as in Study 1, which would be an extremely rare event to occur by chance. There are 5 conditions and thus 5! = 120 possible orderings, giving a probability of < 0.01 of seeing the Study 1 pattern repeated in Study 2. This strengthens our suggestion that the incidence of password recording reflects the difficulty of memorability across the conditions.

5.3 Study 3

The goal of Study 3 was to replicate Study 2 with stronger passwords. We were surprised to find that so many participants in Studies 1 and 2 were able to remember their passwords for the duration of the study. One possible reason for the high memorability was that participants were simply able to remember 21 bits of randomness. To test our hypothesis, we conducted the third study where the theoretical password space of the password systems was increased to 28 bits, giving participants more information to remember (as well as more secure passwords). Study 3 followed the same methodology as Study 2, and participants were recruited from MTurk and completed the entire study online. In all studies, participants from earlier studies were excluded from participation.

For Study 3, the PassTiles passwords were reconfigured to have an 8×10 tile grid, with passwords consisting of 6 tiles. The assigned text passwords consisted of 6 lowercase letters, and the chosen text passwords were restricted to exactly 6 characters.

There were 92 participants (35 female) in Study 3, with a median age of 27. 28 participants were students, and the majority of participants came from either India or the USA.

Memory times for Study 3 were very similar to those in the earlier studies, and similar across the study conditions (Table 9). Although we expected that the increased password strength would lead to participants being able to remember their passwords for less time, it appeared that most participants were still able to remember their passwords for the duration of the study. Once again, we found no significant differences in memory time between Assigned Text and any of the graphical password conditions, or among the three graphical password conditions.

Table 9: Descriptive statistics for memory time (in hours)(Study 3).

	Mean	SD	Median	Skewness	Kurtosis
BPT	169.29	53.55	171.85	-2.52	8.94
IPT	153.14	43.95	161.75	-0.60	1.16
OPT	152.12	53.72	168.70	-1.35	3.72
AST	163.99	33.45	170.95	-1.68	3.67
CHT	175.46	15.69	170.78	-0.02	0.24

Surprisingly, having more secure passwords did not appear to cause participants to reset their passwords more often than in Study 2 (Table 10). As in Studies 1 and 2, we found no significant differences in the number of password resets between Assigned Text and each of the graphical password conditions, nor were the differences between the graphical password conditions significant.

Login times for the graphical password conditions were higher in Study 3 than in Studies 1 and 2 (Table 11). This

Table 10: Descriptive statistics for password resets (Study 3).

	Mean	SD	Median	Skewness	Kurtosis
BPT	0.93	3.20	0	3.70	13.76
IPT	1.42	1.61	0	0.60	-0.97
OPT	1.63	2.87	0	2.87	9.75
AST	1.27	1.87	0	1.37	1.28
CHT	0.15	0.67	0	4.47	20.00

Table 11: Login time statistics (seconds) (Study 3).

	Mean	SD	Median	Skewness	Kurtosis
BPT	33.48	14.44	33.64	0.11	-0.45
IPT	33.00	10.21	32.40	0.65	0.80
OPT	58.36	20.92	50.92	0.97	0.91
AST	11.48	9.58	6.42	1.36	0.76
CHT	7.05	3.90	6.25	3.26	12.51

Table 12: Pairwise Wilcoxon tests of login times using Bonferroni adjustment (Study 3).

	U	p
BPT vs. IPT	179.00	1.000
BPT vs. OPT	42.00	0.006
BPT vs. AST	189.50	0.002
BPT vs. CHT	277.00	< 0.001
IPT vs. OPT	47.00	< 0.001
IPT vs. AST	335.00	< 0.001
IPT vs. CHT	479.00	< 0.001
OPT vs. AST	284.00	< 0.001
OPT vs. CHT	380.00	< 0.001
AST vs. CHT	176.50	1.000

Table 13: Descriptive statistics for memory time (in hours) (Study 4).

	Mean	SD	Median	Skewness	Kurtosis
IPT	13.22	15.85	16.10	1.27	1.65
OPT	22.14	26.42	0.24	0.59	-1.38
AST	6.68	16.69	0.04	2.22	3.27

increase is probably due to having to locate and click on 6 tiles (rather than 5) and the enlarged grid size. Interestingly, login times for the text password conditions remained comparable to the times seen in Studies 1 and 2 (although the passwords were longer). A Kruskal-Wallis test showed significant differences in login time ($\chi^2(3) = 42.78, p < 0.001$), and post-hoc pairwise Wilcoxon tests (Table 12) showed that almost all differences were significant, and the only insignificant differences were between Blank PassTiles and Image PassTiles, and between Assigned Text and Chosen Text. The lack of difference between Blank PassTiles and Image PassTiles appears consistent across the three studies, and is likely due to the very similar nature of the task.

Figure 6(c) shows the reported frequency of password recording by condition. Although the numbers varied, we saw the same pattern of highest to lowest frequency that was seen in Studies 1 and 2, supporting our added hypothesis. A very high proportion of participants in Assigned Text reported writing their passwords down, and this probably accounts for the unexpectedly high memorability and low login times seen in the results.

5.4 Study 4

The goal of Study 4 was to test our hypotheses while reducing the likelihood of password recording. After conducting Studies 1 to 3, it appeared that differences in memorability were being obscured by participants writing their passwords down, and using the recorded passwords to aid their memories. While this is a legitimate technique for coping with difficult memory tasks, it makes it difficult to gain a deeper understanding of how memory for passwords works. In addition, writing passwords down can be a security risk, and we were unhappy that participants felt unable to remember their passwords without writing them down.

Study 4 was designed to investigate the same questions about how retrieval types affect password memorability, but we made a few modifications to the procedure that we hoped would discourage participants from writing their passwords down. We removed session 2 (where participants received email requests to complete tasks on the websites), and when participants completed session 1, we did not immediately tell them that they would be asked to complete session 3, and were vague about the possibility of future tasks. We reasoned that if participants did not think they would need to know their passwords in future, they would not be as likely to write them down. Since factors such as intent and motivation also affect memorability [24], we decreased the total duration of the study to two days to make the memory task easier.

Study 4 used the original 21 bit passwords, and was completed entirely online using participants from MTurk. Apart from the description of the study duration, the instructions were exactly the same as those in Studies 2 and 3. Since part of our goal was to design a feasible graphical password scheme, we decided to only include the schemes showing the most promising usability and security: Image PassTiles, Object PassTiles and assigned text. There were 57 participants (23 female) in Study 4.

Figure 6(d) shows the frequency of reported password recording in Study 4. The modifications to the procedure appear to have had the desired effect, since a lower proportion of participants reported writing their passwords down than in Studies 1, 2, or 3. A chi-squared test of differences in password recording between conditions in Study 4 showed no significant differences in password recording between the three study conditions.

As in the earlier studies, we examined memory time (Table 13) and resets (Table 14) as a measure of the memorability of the passwords. We hypothesized that there would be significant differences in memorability between the Image PassTiles, Object PassTiles and Assigned Text conditions. A Kruskal-Wallis test of memory time showed significant differences between conditions ($\chi^2(2) = 8.52, p = 0.014$), and post-hoc Wilcoxon tests using a Bonferroni adjustment showed that participants were able to remember Object PassTiles passwords significantly longer than Assigned Text ($U = 272.00, p = 0.005$). A Kruskal-Wallis test of resets showed no significant differences between conditions.

Similar to Studies 1, 2, and 3, login times were not normally distributed in Study 4, and were longer for Object PassTiles (Table 15). We hypothesized that there would be

Table 14: Descriptive statistics for password resets (Study 4).

	Mean	SD	Median	Skewness	Kurtosis
IPT	0.89	1.05	0	0.55	-1.33
OPT	0.65	1.22	0	1.94	2.99
AST	0.90	1.37	0	0.95	-1.15

Table 15: Descriptive statistics for login times (Study 4).

	Mean	SD	Median	Skewness	Kurtosis
IPT	14.30	9.07	11.50	2.18	5.55
OPT	43.65	13.35	44.22	0.02	0.88
AST	6.59	1.82	7.00	0.17	-0.53

significant differences in login time between the three conditions, and a Kruskal-Wallis test of login time showed significance ($\chi^2(2) = 20.86, p < 0.001$). Post-hoc pairwise comparisons using a Bonferroni adjustment showed that all pairwise differences were significant (Table 16). These results suggest that the memorability of recognition-based passwords is superior to that of recall-based passwords.

Table 16: Pairwise Wilcoxon tests of login times using Bonferroni adjustment (Study 4).

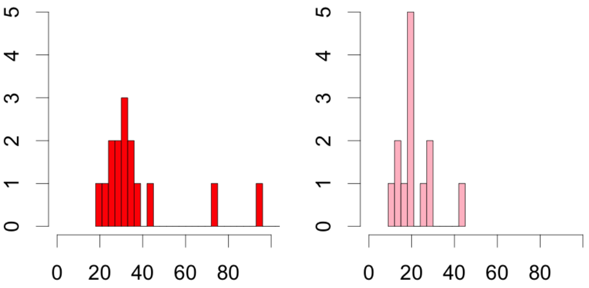
	U	p
Image Passtiles vs. Object Passtiles	3.00	< 0.001
Image Passtiles vs. Assigned Text	81.00	0.011
Object Passtiles vs. Assigned Text	81.00	0.001

5.5 Study 5

The goal of Study 5 was to test our hypotheses with a new condition, based on a modification of Object PassTiles. In Studies 1 to 4, it was apparent that the Object PassTiles condition had good memorability and high learnability, but its usability was severely limited by long login times. We hypothesized that the long login times were due to the shuffling feature, which put the password object images in a different position at every login, causing the user to have to spend time searching for the correct tiles. Since password tile entry order did not affect login success, we were able to gather data about the order in which participants clicked on their password tiles. We analyzed the spatial position of these tile clicks, and graphed them as heatmaps, which can be seen in Figure 8.

Figures 8(a) and 8(b) show the heatmaps for Blank PassTiles and Image PassTiles. In these, a clear pattern can be seen from the top left to the bottom right, indicating that participants took advantage of the spatial position of their password tiles to find and click them more efficiently. However, in Object PassTiles (Figure 8(c)), the pattern is less clear, and we speculate that this is due to participants having to search the whole screen for each object image. Since the object images are shuffled at every entry, participants were not able to rely on recall memory for the locations of their object images. Such a process would be time-consuming and would explain the long entry times seen in the Object PassTiles data.

In our earlier studies, we included the shuffling feature because it forced users to rely on recognition memory, rather than recall of where the tiles were located. In deployed recognition-based graphical password systems (i.e., Passfaces), the shuffling feature has been included as a defence against shoulder-surfing attacks, but it cannot defend against capture attacks that record the password entry attempt. In an effort to decrease the login time for Object PassTiles passwords, we created a new variant of PassTiles: No-Shuffle Object PassTiles (NPT). No-shuffle PassTiles works exactly like Object PassTiles, but the object images remain in the same password tiles at every login, allowing users to leverage recall memory when locating their password tiles.



(a) OPT

(b) NPT

Figure 7: Distributions of login time (in seconds) by number of participants for each study condition (Study 5).

Table 17: Descriptive statistics for login times. (Study 5)

	Mean	SD	Median	Skewness	Kurtosis
OPT	42.23	27.69	31.67	2.00	3.27
NPT	21.30	8.77	20.00	1.51	3.30

In Study 5, we used the same basic study procedure to compare Object PassTiles with No-shuffle PassTiles. The study was conducted online, using MTurk as a source of participants. There were 29 participants (10 female). We were primarily interested in the usability of the system and the associated login times. We hypothesized that login times would be significantly longer for Object PassTiles than No-shuffle PassTiles. Figure 7 shows the distribution of login times in Study 5 and table. The distributions of login time were shown to be approximately normal, and we conducted a t-test comparing the login times for No-shuffle PassTiles and Object PassTiles. We found that the login time for No-shuffle PassTiles was significantly shorter than for Object PassTiles ($t(19) = 2.85, p = 0.005$). This result indicates that by combining recall and recognition memory, we may be able to gain good memorability with acceptable login times.

6. DISCUSSION

The goal of our studies 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. Our studies showed that users were able to remember recognition-based graphical passwords better than recall-based graphical passwords. Studies 1 and 2 did not find significant differences in memorability, and when we increased the password space in Study 3, we were surprised to find that there were still few differences in memorability. It appeared that many participants were writing their passwords down, which was obscuring differences in memorability between the study conditions. When we looked at the frequency of reported password recording in Studies 1, 2 and 3, we found that fewer people reported writing their passwords down in the recognition condition (Object PassTiles), and the most people reported writing them down in the free recall conditions (Blank PassTiles and Assigned Text). We suggest that this constitutes evidence to support the hypothesis that recognition-based graphical passwords are more memorable. In Study 4, we modified the study procedure to avoid having participants write their passwords down, and we found that participants remembered Object PassTiles passwords significantly longer than assigned text passwords.

Although our studies showed that leveraging recognition memory in graphical password systems did help users to remember their passwords, the associated login times were very slow. The average login time for Object PassTiles was around 30 seconds, which is too slow for widespread use. In the free recall and cued-recall conditions, participants had a harder time remembering their passwords, but when they did remember them, they were able to log in faster. Although the literature on memory retrieval predicted success for recognition-based graphical passwords, it did not predict the increased login times seen (largely because the topic of speed is not usually addressed in the memory literature).

In Study 5, we modified Object PassTiles to remove the shuffle mechanism, and we hoped that this would allow users to leverage recognition or recall memory. We found that login times for No-Shuffle Object PassTiles were significantly shorter than those in Object PassTiles. However, the average login time was approximately 20 seconds, which was still longer than average login times for Image PassTiles. We interpreted this to mean that participants were using a combination of recall and recognition memory when recalling their passwords.

The higher login times in the recognition condition stem from differences in the retrieval processes. Recognition memory involves making a binary decision for each image while traversing the entire image set. Using this decision-making process to recognize an entire password can be very slow. In contrast, recall memory involves fewer but more complex tasks. The user is less likely to successfully complete these tasks, but when successful, the process is faster. Figure 8 in the appendix shows heatmap diagrams of where participants clicked on their n^{th} password tiles for PassTiles passwords. The heatmaps show a top-down, left-to-right pattern of where participants clicked on their tiles. This seems to point to the inefficiency of the search process in recognition-based graphical passwords. Since users are not able to anticipate which area of the grid to search, they resort to a time-consuming tile-by-tile approach to search for their password images.

Recognition-based graphical passwords have good memorability, but they do require some care in deployment. If password systems are designed to leverage recognition memory, then multiple password interference becomes a large issue. Care must be taken to ensure that image sets do not overlap, since it could be problematic if users recognized password images on the wrong website. (Our study used non-overlapping image sets.) Interference might also become an issue when resetting passwords, and it would be ideal if reset passwords had no overlapping images. However, gathering and storing large image sets adds extra work to creating and maintaining an authentication system.

The work presented here does not address the threat of shoulder surfing, but it is worth considering that the passwords used in these studies could be easily shoulder-surfed. The shuffling discourages casual shoulder-surfing of Object PassTiles passwords, but does not protect the system against attacks that use a camera to record password entry. As a compromise between situations where shuffling could protect against the threat of shoulder-surfing, and the increased login times caused by shuffling, we speculate that a system like No-Shuffle Object PassTiles could include a shuffle button. This button could be used in situations where other people were present, but could be ignored when there was no threat of shoulder-surfing.

Our studies showed that leveraging recognition in graphical passwords leads to good memorability, but other considerations such as login times and pragmatic concerns need to be taken into account when choosing recognition-based passwords. Different methods of memory retrieval bring different advantages and disadvantages to the memorability and usability of graphical passwords, and by designing with these affordances in mind, we can best leverage users' ability to remember their passwords. Future work on this topic might now focus on field studies of schemes like NPT, deployed in actual practice.

7. CONCLUSION

Graphical passwords have been proposed as an alternative to text passwords that may have superior memorability. Three categories of graphical passwords have been distinguished on the basis of memory retrieval, but which is best has been an open question. The studies presented here investigated how the different forms of memory retrieval affect the memorability of assigned passwords. Assigned text passwords were compared to the three different kinds of assigned graphical passwords, each leveraging a different kind of retrieval: recall, cued-recall, or recognition.

The results of the studies showed that cued-recall was better than free-recall, and that recognition-based graphical passwords were more memorable than recall-based passwords. However, login times were slower with recognition memory. When we modified PassTiles to allow users to take advantage of *both* recall and recognition memory, we found that memorability was good and login times were faster. Pragmatic issues such as password interference need to be considered, but the results of our studies constitute an answer to the open question of memory retrieval and graphical password design.

8. ACKNOWLEDGEMENTS

The first author acknowledges an NSERC Canada Graduate Scholarship, and the second author acknowledges an NSERC Discovery Grant. We also acknowledge funding from NSERC ISSNet.

9. REFERENCES

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

[2] R. Biddle, S. Chiasson, and P. C. van Oorschot. Graphical Passwords: Learning from the First Twelve Years. *ACM Computing Surveys*, 44(4), 2012.

[3] S. Chiasson, C. Deschamps, E. Stobert, M. Hlywa, B. Freitas Machado, A. Forget, N. Wright, G. Chan, and R. Biddle. The MVP Web-based Authentication Framework. In *Financial Cryptography*, 2012.

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

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

[6] D. Davis, F. Monrose, and M. K. Reiter. On user choice in graphical password schemes. In *Proceedings of the 13th USENIX Security Symposium*, Berkeley, CA, USA, 2004. USENIX Association.

[7] A. De Angeli, L. Coventry, G. Johnson, and K. Renaud. Is a picture really worth a thousand words? Exploring the feasibility of graphical authentication systems. *International Journal of Human-Computer Studies*, 63:128–152, 2005.

[8] S. Designer. John the Ripper Password Cracker, August 2011. `http://www.openwall.com/john`.

[9] H. C. Ellis and R. R. Hunt. *Fundamentals of Human Memory and Cognition*. Wm. C. Brown Publishers, Dubuque, Iowa, 4th edition, 1989.

[10] D. Florencio and C. Herley. A large-scale study of web password habits. In *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, New York, USA, 2007. ACM.

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

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

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

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

[15] P. G. Kelley. 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, 2010. ACM.

[16] A. Kittur, E. H. Chi, and B. Suh. 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, 2008. ACM.

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

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

[19] D. L. Nelson, V. S. Reed, and C. L. McEvoy. 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, September 1977.

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

[21] A. Paivio. *Imagery and Verbal Processes*. Holt, Rinehart, and Winston, 1971.

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

[23] Real User Corporation. The Science Behind Passfaces. Technical report, Real User Corporation, June 2004.

[24] B. Schwartz, A. Benjamin, and R. Bjork. The inferential and experiential bases of metamemory. *Current Directions in Psychological Science*, 6(5):132–137, 1997.

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

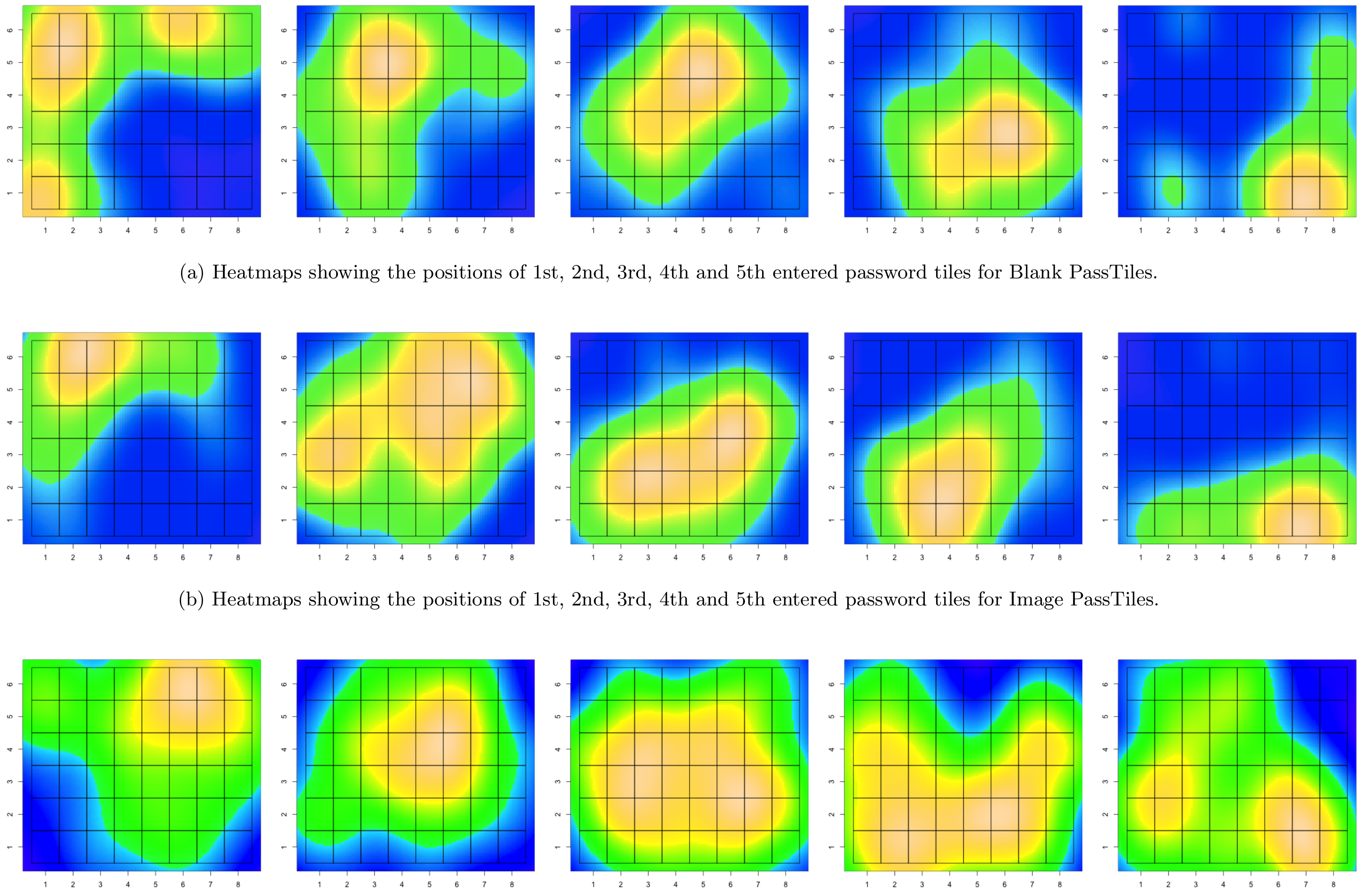
[26] P. C. van Oorschot and J. Thorpe. On Predictive Models and User-Drawn Graphical Passwords. *ACM Transactions on Information and System Security*, 10(4), January 2008.

[27] P. C. van Oorschot and J. Thorpe. Exploiting predictability in click-based graphical passwords. *Journal of Computer Security*, 19(4):669–702, 2011.

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

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

APPENDIX



(a) Heatmaps showing the positions of 1st, 2nd, 3rd, 4th and 5th entered password tiles for Blank PassTiles.

(b) Heatmaps showing the positions of 1st, 2nd, 3rd, 4th and 5th entered password tiles for Image PassTiles.

(c) Heatmaps showing the positions of 1st, 2nd, 3rd, 4th and 5th entered password tiles for Object PassTiles.

Figure 8: Heatmaps showing the spatial positioning of clickpoint entries in PassTiles. A top-down, left-to-right pattern is seen in all three variants of PassTiles.