

Authentication Using Graphical Passwords: Effects of Tolerance and Image Choice

Susan Wiedenbeck
Jim Waters
College of IST
Drexel University
Philadelphia, PA 19104
+1 215 895-2490
sw53@drexel.edu
jw65@drexel.edu

Jean-Camille Birget
Computer Science Department
Rutgers University
Camden, NJ
+1 856 225-6653
birget@camden.rutgers.edu

Alex Brodskiy
Nasir Memon
Computer Science Department
Polytechnic University
Brooklyn, NY 11201
+1 718 260-3970
abrods01@utopia.poly.edu
memon@poly.edu

ABSTRACT

Graphical passwords are an alternative to alphanumeric passwords in which users click on images to authenticate themselves rather than type alphanumeric strings. We have developed one such system, called PassPoints, and evaluated it with human users. The results of the evaluation were promising with respect to memorability of the graphical password. In this study we expand our human factors testing by studying the effect of tolerance, or margin of error, in clicking on the password point and the effect of the image used in the password system. In our tolerance study, results show that accurate memory for the password is strongly reduced when using a small tolerance (10 x 10 pixels) around the user's password points. This may occur because users fail to encode the password points in memory in the precise manner that is necessary to remember the password over a lapse of time. In our image study we compared user performance on four everyday images. The results indicate that there were few significant differences in performance of the images. This preliminary result suggests that many images may support memorability in graphical password systems.

Categories and Subject Descriptors

H.5.2 [Interfaces and Representation]: User Interfaces – Graphical user interfaces; K.6.5 [Computing Milieux]: Security and Protection – Authentication.

General Terms

Security, Human Factors, Design, Experimentation.

Keywords

Graphical passwords, authentication, password images, tolerance, PassPoints, password security, human factors, usable security.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SOUPS'05, July 6–8, 2005, Pittsburgh, PA, USA.

Copyright 2004 ACM 1-58113-000-0/00/0004...\$5.00.

1. INTRODUCTION

Because of increasing threats to networked computer systems, there is great need for security innovations. Security practitioners and researchers have made strides in protecting systems and, correspondingly, individual users' digital assets. However, the problem arises that, until recently, security was treated wholly as a technical problem – the system user was not factored into the equation. Users interact with security technologies either passively or actively. For passive use understandability may be sufficient for users. For active use people need much more from their security solutions: ease of use, memorability, efficiency, effectiveness and satisfaction. Today there is an increasing recognition that security issues are also fundamentally human-computer interaction issues [15, 25].

Authentication is the process of determining whether a user should be allowed access to a particular system or resource. It is a critical area of security research and practice. Alphanumeric passwords are used widely for authentication, but other methods are also available today, including biometrics and smart cards [11, 19]. However, there are problems of these alternative technologies. Biometrics raise privacy concerns and smart cards usually need a PIN because cards can be lost. As a result, passwords are still dominant and are expected to continue to remain so for some time [10].

Yet traditional alphanumeric passwords have drawbacks from a usability standpoint, and these usability problems tend to translate directly into security problems. That is, users who fail to choose and handle passwords securely open holes that attackers can exploit [9, 14, 16, 20, 22, 29]. The “password problem,” as formulated by Birget in [33], arises because passwords are expected to comply with two conflicting requirements, namely:

1. Passwords should be easy to remember, and the user authentication protocol should be executable quickly and easily by humans.
2. Passwords should be secure, i.e., they should look random and should be hard to guess; they should be changed frequently, and should be different on different accounts of the same user; they should not be written down or stored in plain text.

Meeting these conflicting requirements is almost impossible for humans, with the result that users compensate by creating weak passwords and handling them in an insecure way.

Many problems that users have with alphanumeric passwords are related to memorability of secure passwords. In an attempt to create more memorable passwords, graphical password systems have been devised. In these systems authentication is based on clicking on images rather than typing alphanumeric strings. Several kinds of graphical passwords have been invented. In recent work we have created a new kind of graphical password system, called PassPoints, and have done studies of its human factors characteristics compared to alphanumeric password [33]. In this paper we report on further research on usability and memorability of our system under different conditions. In specific we investigate the effect of the tolerance, or the margin of error, allowed when entering one's password points and the effect of the choice of images used in the password system.

The following section briefly describes the difficulties users have with traditional passwords and the alternative of graphical passwords. This is followed by a description of PassPoints and a summary of our recent results comparing PassPoints to alphanumeric passwords. Section 3 reports on the tolerance study and Section 4 the images study. This is followed by the conclusion in Section 5.

2. BACKGROUND TO THE RESEARCH

2.1 Users' Problems with Passwords

Users' propensity to handle alphanumeric passwords insecurely arises largely from long-term memory (LTM) limitations. Users have difficulty remembering complex, pseudo-random passwords over time. The Power Law of Forgetting [2] describes rapid forgetting soon after learning, followed by very slow decay over the long-term. Psychological theories have identified decay over time and interference with other information in LTM as underlying reasons for forgetting [34]. A user is likely to forget a password that is not used regularly, as the memory is not "refreshed" or "activated" sufficiently often. When users have multiple passwords, today practically a universal condition, interference becomes a possibility. The user may either jumble the elements of the different passwords or remember the password but confuse which system it corresponds to.

Users normally cope with password memory problems by decreasing the complexity and number of passwords, thereby reducing password security. A secure password should be 8 characters or longer, random, with upper-case characters, lower-case characters, digits, and special characters. Such passwords lack meaningful content and can be learned only by rote memorization, a weak way of remembering [28]. Generally, users ignore such password recommendations, using instead short, simple passwords that are relatively easy to discover using dictionary attacks or attacks based on knowledge of the user. Recent surveys have shown that users often choose, short, alphabetic-only passwords consisting of personal names of family or friends, names of pets, and even the word "password" [9, 29]. Users typically write down their passwords, share passwords with others, and use the same password for multiple systems, sometimes with a single digit added on the end [1, 29]. While poor password practices may be largely attributed to memory

problems, there are other factors as well. Some users are naïve about the power of a modern dictionary attack or about the scope of the damage that may occur if their computer is breached. Even if users are somewhat knowledgeable about security, their motivations may get in the way of good practices; they want to get real work done and therefore view authentication as an enabling task that should be gotten over with as quickly as possible [1].

2.2 Graphical Passwords

2.2.1 *Why Graphical Passwords May Be Better*

Most graphical password systems are based on either recognition or cued recall. In recognition-based systems the user must recognize previously chosen images from a larger group of distractor images. The decision is binary: either the image is known (recognized) or not known. In cued recall password systems users must click on several previously chosen areas in an image, cued by viewing the image.

Both types of systems may have memory advantages over alphanumeric passwords. Alphanumeric passwords are based on pure recall (presuming the user has not written the password down). It is known that recognition memory is better than unaided recall [24]. Furthermore, psychological studies show that images are recognized with very high accuracy (up to 98 percent) after a two hour delay, which is much higher than accuracy for words and sentences [30]. In addition, it has been found that error in recognition of images is only 17 percent after viewing 10,000 pictures [31]. Studies of recall also confirm that pictures are recalled better than words [26] and this has led to the tag "picture superiority effect" [23].

Cued recall, as used in graphical password systems, seems to be intermediary between recognition and pure recall. The decision is not binary based on recognition of the image as a whole. The user has to recall his or her click areas within the image. But scanning the image helps the user identify the correct areas.

Other psychological research on images has shown that people can remember detailed visual information in natural scenes [Hollingsworth & Henderson] and that the content, affect, and organization of images influence the ability to remember an image [8, 21]. In terms of choice of memorable images, psychologists have found that coherent images are more memorable than jumbled ones [3]. Also, LTM stores the meaning of an image, not a replica of it [21]; therefore, concrete scenes are likely to be remembered well because of their semantically meaningful content, as opposed to abstract images.

2.2.2 *Graphical Password Systems*

There are several graphical password systems based on recognition [13, 14, 27, 32]. For example, Passfaces [27] worked as follows in Brostoff and Sasse's empirical study [10]. To create a password the user chooses four images of human faces from a large portfolio of faces. When logging in, the user sees a 3x3 grid with nine faces, consisting of one face previously chosen by the user and eight decoy faces. The user must recognize and click anywhere on the previously chosen face. This procedure is repeated with different target and decoy faces, for a total of four rounds. Only if the user chooses all four correct faces, will he or she successfully log in. Empirical evidence from a field trial [10] shows that Passfaces may be more memorable than alphanumeric passwords. Evidence from another similar system, Déjà Vu [14],

suggest that initially choosing the images from the portfolio is a rather slow process, but the images are easier to remember over time. However, the drawback of all such passwords based on image recognition is that only a small number of images can be displayed, e.g., nine images, one of which is a chosen image. Therefore, an attacker has a 1-in-9 chance of guessing the image. To reduce that chance the login process uses several rounds of recognition. To obtain security similar to that of 8-character alphanumeric password (over an alphabet of 64 characters), 15 or 16 rounds with 9 faces each would be required. This could make the login slow and tedious. Also, using faces as the images has been shown to lead to passwords with very low entropy because people choose faces in predictable ways [12].

Graphical passwords based on cued recall were first discussed by Blonder [*]. In such a scheme the user chooses several locations in an image to create a password. To log in the user must click on or close to those locations. There are no multiple rounds of images, just a single image. In an implementation of this scheme [7] the image had predefined click objects or regions that were outlined by thick boundaries. The users chose the password from these objects and logged in using them (although thick boundaries were not visible when logging in). A click anywhere within the boundary was considered correct. A problem with this scheme was that the number of predefined click regions was relatively small so the password had to be quite long to be secure (e.g., 12 clicks). Also, since they had to outline readily identifiable objects or regions, they had to use rather simple, artificial images instead of complex, real-world scenes.

Our system, PassPoints [33], is based on Blonder's idea of representing the password as multiple clicks on a single image. However, it overcomes some of the limitations of his scheme. There are no artificial predefined boundaries around areas of the image. This means that users may choose any place in the image as a password click point. When a click point is chosen the system automatically hashes the password for security and calculates a tolerance around the chosen pixel that becomes the user's tolerance [4]. The size of this tolerance may be varied, but for a large password space should not be too large, e.g., .25 to .50 cm around the actual click point. To log in users must click within the tolerance of their chosen click points. Their memory is cued by the image as they enter their password. Without artificial predefined boundaries, more intricate images, such as natural scenes, can be used. An intricate image has hundreds of memorable points, so fewer clicks are needed to make a password hard to guess, for example 5-6 click points can make more passwords than an 8-character Unix-style password over a standard 64-character alphabet. Similarly, recognition-based passwords (Passfaces, etc.) would have to have many rounds to equal the password space of PassPoints using 5-6 clicks. The system or the user could provide the image. The main requirement is that it be a complex image that is visually rich enough to have many potentially memorable click points.

We compared PassPoints to alphanumeric passwords in a laboratory study [33]. Our main interest was to evaluate the learning and memorability of our graphical passwords. There were 40 participants, and half were assigned to each group. Participants created and practiced either an alphanumeric or a graphical password. The participants subsequently carried out three longitudinal trials to input their password over the course of five weeks. The results showed that the graphical password group

created a valid password with fewer difficulties than the alphanumeric group. However, the graphical group took longer and made more errors in carrying out the practice. This was expected given that the graphical group was using a kind of password that was entirely new to them. More importantly, all graphical users were able to reach the learning criterion within several minutes. In longitudinal trials the two groups performed similarly on memory of their password over five weeks. User perceptions of the two password systems, collected by questionnaire, were quite similar.

2.2.3 Research Questions

Given largely encouraging results of our empirical testing, we have proceeded to carry out empirical studies of some of the key parameters in the PassPoints system: the effect of the tolerance around user click points and the effect of image content on user performance.

Varying the tolerance may affect both memory accuracy and motor activities; for small margins of error users may need to have a more accurate memory of their click points, as well as more attention to motor skill in clicking [17]. Yet the trade-off is that small tolerances increase the space of possible passwords and therefore make passwords more secure (less guessable).

Likewise, the nature of the images used in the system may have a large effect on people's ability to remember their click points. There is a trade-off here, too, because allowing users to choose their own images, which would be quite possible to do in this system, may lead to high memorability for an individual (e.g., a family photo), but at the same time may result in images with poor security characteristics (e.g., few memorable click points, images that are guessable with knowledge of that user). These issues are important because successful performance with a graphical password system may hinge critically on the trade-offs. Therefore, our objective was to vary these conditions within reasonable bounds and evaluate how they affect performance.

3. TOLERANCE STUDY



The objective of this study was to understand the effect of different tolerance sizes around user click points. The tolerance can be varied in the system. Our question is how does varying the tolerance affect success in graphical password use. In our previous experiment we used a tolerance of 20 x 20 pixels (.53 cm²). However, this relatively large tolerance restricted the password space more than we liked. Therefore, we experimented with smaller tolerances to see how they affect user performance.

3.1 Methodology

Thirty-two undergraduate students, ranging from their first year to their last year of studies, participated in the experiment. Ten were female and 22 were male. The mean age of participants was 22.7 (SD=1.33). Most of the participants were majoring in information systems. They all used PCs frequently.

The PassPoints system used in this study was the same as in [33], except that it used a different image. The interface included the image used for testing and several buttons. A single image used in this experiment depicted a colorful scene of children painting murals in a room. The size of the image was 451 x 331 pixels. Two tolerance were around the click points were used: 14 x 14 pixels, and 10 x 10 (Table 1).

Table 1. Tolerance around click points

Tolerance	Size in cm ²	Example
10 x 10	.26cm ²	
14 x 14	.37cm ²	

The image occupied about half of the full screen (Figure 1). The rest of the screen consisted of a background with five buttons and with empty space to present instructions in certain phases of the experiment. The Submit button was used to submit the password when the user had entered the points. The Undo and Clear buttons were used to correct a password before it was submitted; the Clear button erased all password points input so far; the Undo button erased only the user's most recent point. The See My Password button allowed the user to view his or her password during the learning phase and under certain circumstances in the following retention phase. The Quit button allowed a user to quit the experiment. This button was placed to the side to avoid accidental quitting. All instructions for the participants were given on the screen and feedback on correctness of a password input was given on screen after the user clicked the Submit button. The online testing system also included a questionnaire that asked the user's perceptions of the password system. The questions were answered on a 7-point Likert-type scale ranging from strongly agree (1) to strongly disagree (7).



Figure 1. Image used in tolerance study.

The study was carried out in three sessions in a closed laboratory that seated up to 25 people. Each participant was sat at a PentiumIV computer with a high resolution 19 inch monitor. The participants were randomly assigned to the two tolerance conditions. There were 16 participants in both the 14 x 14 and the 10 x 10 groups. The session began with an approximately 7 minute Powerpoint presentation that introduced the experiment and explained the concept of graphical passwords. In the experiment, which lasted about 30 minutes, the participants first entered demographic data. Moving on to the experiment, instructions on the screen guided the participants to create a valid password consisting of 5 click points, none of which were within the tolerance around another click point. They were told that they would have to remember the points and the order in which they were input. A graphical password of 5 points was used based on an analysis which showed that in terms of security 5 click points provide a password space as large as or larger than an

alphanumeric password of 8 characters [33]. When the participant had created a valid password, the password was displayed as feedback to the participant about the locations of the click points and the size of the tolerance. The display showed the image with a heavy outline of the size of the tolerance around each point chosen. The points were also numbered 1 to 5 to indicate their order of input (Figure 2).



Figure 2. Excerpt of image with tolerance and click order

When the participant had created a valid password, the learning phase began. To reinforce the password the participant entered the password repeatedly until he or she achieved ten correct password inputs. Participants received binary feedback on the correctness of each password input and could see an on-screen count of how many correct and incorrect entries they had made. If the user was not able to remember the password, he or she could click on Show My Password, which displayed the password image with the password points indicated. After the learning phase, the participant filled out the questionnaire online. This was designed to gather user perceptions and act as a distractor between the learning phase and the first retention trial.

In the retention phase password retention was measured at the end of the first session (R1) and one week later (R2). The participant had to enter the password correctly one time. The retention trials took 5 minutes or less. The trial was over as soon as the participant entered the correct password. If the participant entered an incorrect password, the system gave feedback that the password was wrong, and the participant was instructed to re-enter the password. If the user failed to input the password correctly after five attempts, the Show My Password button was enabled and the participant could view the password, then make another attempt to input it.

3.2 Results

We recorded results about the three phases of the study: password creation, learning, and retention, as discussed below.

Two participants in the 10 x 10 group created an invalid passwords and had to try again. There were no password creation errors in the 14 x 14 group. The errors were not serious. An example of an error is that the participant entered the wrong number of points (e.g., 4 rather than 5) out of inattention to the on-screen instructions. There were no significant differences in the number of attempts or the time to create a valid password.

In the learning session participants entered their password repeatedly until they had accomplished 10 correct inputs. We measured the number of attempts to meet the criterion and the time. The means and standard deviations are show in Table 2.

T-tests were used for the analyses. The t-test for number of incorrect submissions was not significant. The time for incorrect

submissions was marginal, $t(31)=3.46$, $p<.071$. The time for correct submission was not significant. Further results showed that the two groups were almost equivalent in terms of the number

Table 2. Means (SD) in learning phase

	Tolerance	
	14 x 14	10 x 10
Number of incorrect submissions	1.56 (3.65)	4.81 (6.83)
Time for incorrect submissions (sec)	116.08 (78.52)	181.43 (115.50)
Time for correct submissions (sec)	10.52 (4.35)	12.83 (4.01)

of individuals who entered their password 10 times without any errors – 6 in the 10 x 10 group and 7 in the 14 x 14 group. However, in the 14 x 14 group 15 out of 16 individuals succeeded with only one or two extra attempts. By contrast, individuals in the 10 x 10 tolerance took many more trials (Table 3).

Table 3. Number of participants making incorrect submissions by tolerance in the learning phase

	Number of Incorrect Submissions												
	0	1	2	3	4	5	6	9	7	8	9	15	25
14 x 14	7	6	2									1	
10 x 10	6	2		1		1	2	1	1			1	1

Tables 4 and 5 show the means and standard deviations of the retention phase. R1 is the password input at the end of the first session, after the questionnaire/distractor task; R2 is the session one week later. Recall that participants had only to enter their password correctly one time in the retention trials.

Table 4. Means (SD) in R1 retention trial

	Tolerance	
	14 x 14	10 x 10
Number of incorrect submissions	0.19 (0.54)	0.31 (0.87)
Time for incorrect submissions (sec)	2.05 (6.65)	1.85 (5.13)
Time for correct submission (sec)	10.01 (8.85)	8.85 (2.94)

Table 5. Means (SD) in R2 retention trial

	Tolerance	
	14 x 14	10 x 10
Number of incorrect submissions	0.94 (2.14)	3.12 (3.20)
Time for incorrect submissions (sec)	11.42 (23.56)	50.88 (65.20)
Time for correct submission (sec)	15.60 (6.97)	16.85 (9.86)

Two-way mixed model ANOVAs were used for the analyses with tolerance as the between-subjects factor and retention trial (R1/R2) as the within-subjects factor. The ANOVA for the number of incorrect submissions showed that the effect of retention trial was significant, $F(1,30)=12.62$, $p<.001$ with a higher number of incorrect attempts in R2. The within-subjects effect of tolerance was also significant, $F(1,30)=5.45$, $p<.027$, with tolerance 10 x 10 taking more attempts. The interaction of retention trial and tolerance was significant, $F(1,30)=4.23$, $p<.049$, (Figure 3). Follow-up using Newman-Keul’s test indicated that the number of incorrect submissions in the 10 x 10 group in trial R2 was significantly higher than any of the other tolerance by trial groups. In terms of individual participants, our data show that in R1 there were two participants in each tolerance that made errors submitting their password. By contrast, in R2 11 participants in the 10 x 10 group made at least one error, while 5 participants in the 14 x 14 group made at least one error. Furthermore, we examined how many participants “failed” to log in by the criterion of making a correct log in within 4 or less attempts. This criterion was chosen because existing password systems often block users if they make repeated errors on input. In fact, our participants were allowed to continue attempting to log in as long as they wished, but after 4 unsuccessful attempts the Show My Password button became active and the participants could look at their password. In the 10 x 10 group 7 of 16 participants (43.75 percent) failed to log in, while in the 14 x 14 group only 2 of 16 failed (12.5 percent). There was a significant difference on failure between the groups $t(30)=2.63$, $p<.015$.

The analysis of time for incorrect submissions showed that there was a main effect of retention trial, $F(1,30)=11.13$, $p<.002$, a main effect of tolerance, $F(1,30)=5.03$, $p<.032$, and a significant interaction, $F(1,30)=5.12$, $p<.031$ (Figure 4). A Newman-Keul’s test showed that, as in the previous case, the only significant difference in the interaction was that the 10 x 10 group in R2 took significantly more time on incorrect attempts. Finally, the analysis of correct submission times showed that there was a significant effect of retention trial, $F(1,30)=15.79$, $p<.0001$, but no significant effect of tolerance or the interaction.

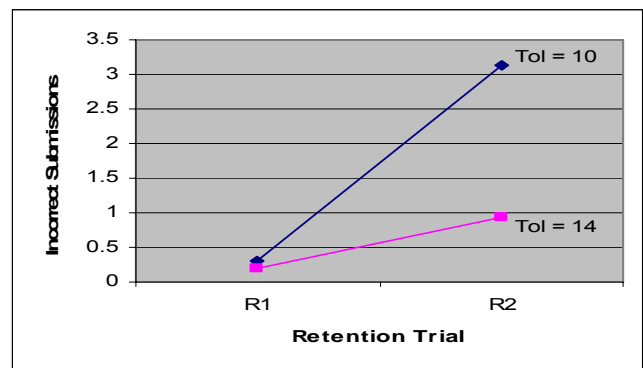


Figure 3. Number of incorrect submissions by tolerance in R1 and R2.

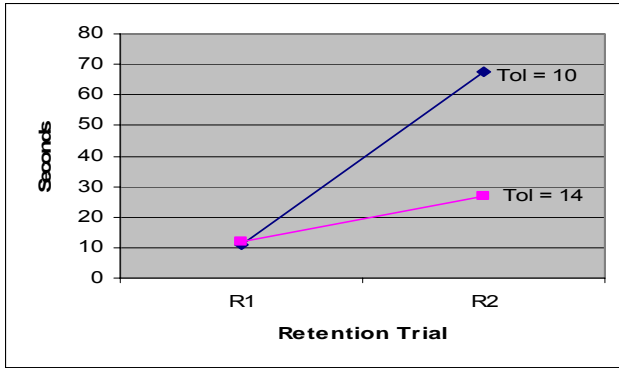


Figure 4. Time for incorrect submissions by tolerance in R1 and R2.

User perceptions measured in a questionnaire found several interesting results. Users were posed the following key questions about graphical password use. Other items in the questionnaire did not target use directly and were meant to increase the time between the learning phase and the R1 retention trial, for example, questions about the understandability of the on-screen instructions or the amount of feedback the user received.

The questions are listed below, and the means and standard deviations are shown in Table 6. Recall that a lower number indicates *higher* agreement with the statement.

1. It did not take me long to input my password correctly 10 times
2. Once I had created my password I was able to input it correctly.
3. Inputting my password was easy.
4. Inputting my password was fast.
5. I think that the password system was pleasant to use.
6. There was too much to remember when using the system.

Table 6. Means (SD) of question responses

	Mean (SD) 14 x 14	Mean (SD) 10 x 10	Mann-Whitney U	p
Q1	1.69 (1.54)	3.31 (2.33)	66.00	.019
Q2	1.44 (1.09)	2.89 (1.82)	67.50	.021
Q3	1.81 (0.98)	3.06 (1.95)	82.00	.086
Q4	2.38 (1.31)	3.31 (2.02)	95.00	.224
Q5	1.81 (0.91)	3.56 (1.86)	51.50	.003
Q6	2.94 (1.65)	3.13 (1.82)	122.50	.838

The questions were analyzed with the Mann-Whitney U test, which is a non-parametric test used for two independent samples. The questions tended to be answered more favorably by the 14 x 14 group in most cases. Exceptions were “inputting my password

was fast,” and “there was too much to remember when using the system.”

3.3 Discussion

Participants had little difficulty creating a valid graphical password, but learning their password via repeated password inputs posed challenges to some. While there were no significant differences, there appeared to be a trend for individuals in the smaller tolerance to make more errors and for the input time for their erroneous password attempts to take longer. Another indicator of this trend is the long trail of participants who took many practice trials in the smaller tolerance (*Fig). The two tolerance groups were essentially equivalent in the number of individuals who input their password 10 times with no errors. In the larger tolerance group 15 of 16 individuals met the criterion of 10 correct inputs with 2 errors or less. In the smaller tolerance group only 8 participants met the criterion with 2 errors or less. The other 8 participants took from 3 to 25 errors inputs before achieving the 10 correct trials. Using a graphical password was a new to the participants and we expected errors in the learning phase, but the long trail of errors in the 10 x 10 group is quite striking. The difficulties that users had in the 10 x 10 group are also reflected in several of the questionnaire items, in which they tend to have poorer perceptions on key items, such as speed of entering their password, ease of using the password, and pleasantness of using the password system. On the other hand, it should be noted that the time for participants in the two groups to input a *correct* password was equivalent.

In the R1 retention trial there were very few incorrect password submissions, 2 in each tolerance group. It should be noted that the mean time for these incorrect submissions was very low. This almost certainly indicates that the errors were slips, in which the participants noticed a slip immediately and submitted it so they could start over. In fact, there was a Clear button provided for this case, but it was seldom used. In the correct submissions the time was essentially equivalent for the two groups. The time of about 9 to 10 seconds is consistent with the results in our prior study of PassPoints [33]. In the earlier study, the time for inputting alphanumeric passwords was about 5 to 6 seconds. We find it encouraging that after a little practice the difference is only a few seconds. A longer input time is expected because of the mouse movement, but it appears that the time is reasonable. Furthermore, the participants were not highly skilled using graphical passwords. We expect users to input graphical passwords faster with continual use and automation of the process.

The results of the R2 retention trial were strikingly different from R1. The main issue was the participants’ ability to remember their graphical password. Participants in the smaller 10 x 10 tolerance made significantly more incorrect submissions than the larger 14 x 14 tolerance. First, it should be noted that there were relatively few errors in the 14 x 14 group in R2, only 15 in total for a mean of slightly less than 1 per person. However, eleven participants had no errors at all, and most of the errors were from two people, one who needed 4 attempts to be successful and the other who needed 8 trials. Only one of the participants took more than 4 trials, our criterion for failure. This is an encouraging result for the 14 x 14 group. The time for these incorrect submissions was also low, suggesting that participants knew where to look for their password points and located and clicked on them quickly. The slightly longer mean time for correct password submissions in this

group may mean that individuals who made an error in their first attempt(s) were slower and more cautious in their subsequent correct attempt. Turning to the 10 x 10 group, we see a strong contrast with the 14 x 14 group. Only 5 participants were able to input their password in the first try without errors, but there was a total of 50 errors among the remaining 11 participants. Seven of them failed according to our criterion of 4 attempts to log in. They also had a much higher mean time to input these incorrect passwords, 51 seconds. This suggests that individuals in the 10 x 10 tolerance had to spend a great deal more time scanning the image to identify their password points. They probably also had to observe the area of their password points very carefully to identify the exact place to click because of the small tolerance. We do not believe that the manual activity of actually clicking in the smaller tolerance is an explanation of the time results, given the results in the learning phase and R1, where the correctness and time to input were very similar between the two groups.

In sum, the 10 x 10 participants were equivalent to the 14 x 14 participants in the first retention, R1, shortly after learning their passwords. However, they performed much more poorly than the 14 x 14 group in the retention phase. We interpret this as largely a matter of precision of memory, not manual precision of clicking. Participants using the 10 x 10 tolerance had to encode their password points more precisely in memory to successfully use the password after a lapse of time. For example, rather than encoding a password point as the “paint brush,” the individual would encode it as the “handle of the paint brush.” While the 14 x 14 group also had to encode their password points relatively precisely in memory, the difference in tolerance gave them a greater margin of error when they had to input the password. We conclude that a tolerance of 14 x 14 pixels can be used successfully by PassPoint users, but a smaller tolerance of 10 x 10 pixels is significantly more difficult, given intermittent use.

4. IMAGES STUDY

The objective of this study was to understand the effect of different images on user performance. Our question is how does varying the image affect success in graphical password use. There is a dearth of knowledge about memorability of specific kinds of images. First, to our knowledge, there is no theory or taxonomy of classes of images that might structure image choice. Second, psychologists have studied images, but much of the research has focused on the memorability of images compared to words and sentences, i.e. the “picture superiority effect.” Studies of characteristics of images exist, but are not highly directive for our purposes. Some research studies have investigated image memorability in the context of free recall of image, others in the context of recognition memory. These studies do not give us sufficient guidance about cued recall of images, as used in Blonder-style systems such as PassPoints. We chose several everyday images based on the existing psychological research and intuition, with the purpose of gaining some initial knowledge about learnability and memorability when using different images. Thus, this is an exploratory study.

4.1 Methodology

Eighty-three undergraduate students participated in the study. There were 42 males and 8 females. The mean age was 22.7 (SD=2.84). Most of the participants were majoring in information

systems. They all used PCs frequently. They did not participate in the tolerance study.

The system set-up was exactly the same as in the tolerance study described above, with two differences. First, only one tolerance was used: 20 x 20 pixels, which equated to square of area .53 cm², as in [33]. Second, four images were used. One image was the children painting murals used in the study above. The other three images represented respectively an indoor swimming pool and its surrounds with people walking around it, a small room with a table holding colorful teapots and crockery, and a city map of the central area of Philadelphia (Figures 1, 5, 6, 7).



Figure 5. Image POOL



Figure 6. Image TEA.



Figure 7. Image MAP.

The study was carried out in four sessions in two closed laboratories, each seating up to 25 people. In each session participants were randomly assigned to the four image conditions. The number of participants in each group was as follows: POOL 20, MURAL 18, TEA 22, and MAP 23.

4.2 Results

In the password creation phase there were 13 errors in which participants failed to create a valid graphical password and had to try again. Except for one person, the participants were able to

make a valid password on the second try. There were no significant differences among the groups in the number of attempts or the time to create a valid password.

The means and standard deviations of the learning phase are shown in Table 7.

Table 7. Means (SD) in learning phase

	Image			
	POOL	MURAL	TEA	MAP
Number incorrect submissions	4.80 (7.16)	1.00 (1.68)	3.23 (5.94)	1.70 (4.57)
Time for incorrect submissions (sec)	160.77 (107.67)	67.58 (48.92)	117.12 (63.12)	113.71 (75.60)
Time for correct submissions (sec)	11.18 (2.46)	8.45 (2.63)	12.14 (15.42)	10.84 (8.29)

Oneway ANOVAs were used for the analyses and Tukey’s HSD for specific comparisons. The ANOVA for the number of incorrect submissions was not significant. The time for correct submission was also non-significant. Finally, time for incorrect submissions was significant, $F(3,79)=2.98$, $p<.036$. Tukey’s HSD showed that the only significant difference was between POOL and MURAL. The table below shows that most participants made 0 to 2 incorrect submissions while practicing. However, there were several individuals in each group who needed many practice trials to meet the criterion of 10 correct password inputs (Table 8).

Table 8. Number of participants making incorrect submissions by image in the learning phase

	Number of Incorrect Submissions														
	0	1	2	3	4	6	7	9	10	13	17	18	20	22	25
POOL	8	4	1	1		1		1			2	1	1		
MURAL	9	5	3				1								
TEA	10	3	2		4				1	1					1
MAP	13	4	3	1	1									1	

Tables 9 and 10 show the means and standard deviations in the retention phase.

Table 9. Means (SD) in R1 retention trial

	Image			
	POOL	MURAL	TEA	MAP
Number incorrect submissions	0.55 (1.57)	0.17 (0.51)	0.14 (0.47)	0.39 (1.31)
Time for incorrect submissions (sec)	7.91 (28.20)	1.18 (3.37)	1.26 (4.40)	3.08 (9.25)
Time for correct submissions (sec)	8.61 (3.27)	7.52 (3.62)	7.43 (2.32)	7.86 (2.03)

A two-way mixed model ANOVA was used for the analyses with image as the between-subjects factor and retention trial (R1/R2) as the within-subjects factor. The ANOVA for the number of incorrect submissions showed that the effect of retention trial was

Table 10. Means (SD) in R2 retention trial

	Image			
	POOL	MURAL	TEA	MAP
Number incorrect submissions	2.75 (3.88)	2.00 (3.33)	2.64 (3.27)	1.30 (3.21)
Time for incorrect submissions (sec)	60.90 (67.35)	27.96 (45.44)	39.99 (52.25)	63.83 (13.31)
Time for correct submissions (sec)	18.50 (5.92)	14.54 (6.58)	13.67 (4.91)	20.29 (15.46)

significant, $F(1,30)=124.40$, $P<.0001$, with a higher number of incorrect attempts in R2. The between-subjects effect of image was not significant, nor was the interaction (*Figure). Only two of the 83 participants made submission errors in the first retention trial, R1. However, 36 of 83 individuals made errors in the second retention trial one week later, and 22 of those individuals made more than the 4 incorrect attempts and thus “failed” (26.5 percent). The number of participants who failed in each group varied from 3 to 8; there were no significant differences by image. The analysis of time for incorrect submissions showed that the only significant effect was retention trial, $F(1,79)=79.00$, $p<.0001$. Finally, the analysis of correct submissions showed that there was again a significant effect of retention trial, $F(1,79)=70.36$, $p<.0001$. There was also a marginal effect of image, $F(3,79)=2.55$, $p<.062$. Tukey’s HSD indicated that performance of the MAP group was lower than the TEA group.

The perceptions of the participants are reported in Table 11. The question numbers refer to the same questions used in the tolerance study. The questions were analyzed with the Kruskal Wallis test, which is a non-parametric test used for k-independent samples. Ryan’s test was used for follow-up testing.

Table 11. Means (SD) of question responses

	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Kruskal Wallis df = 3	p
	POOL	MURAL	TEA	MAP		
Q1	3.40 (2.14)	2.00 (1.33)	2.86 (1.78)	2.09 (1.78)	8.90	.031
Q2	2.60 (2.01)	1.56 (.071)	1.95 (1.23)	1.87 (1.74)	4.97	.174
Q3	2.70 (2.18)	1.44 (0.78)	2.36 (1.25)	1.91 (1.78)	10.25	.017
Q4	3.05 (1.73)	1.89 (1.18)	3.18 (1.62)	2.35 (1.34)	9.73	.021
Q5	2.75 (2.17)	2.22 (1.63)	2.73 (1.64)	2.70 (1.74)	1.81	.595
Q6	2.65 (1.53)	2.28 (1.60)	3.18 (1.89)	2.70 (1.82)	3.37	.338

There were differences on three of the questions. For “it did not take me long to input my password correctly 10 times” the follow-up test showed that the MURAL and MAP groups agreed more strongly with this statement than the POOL group ($p<.05$). For “inputting my password was easy” and “inputting my password was fast” the significant difference ($p<.05$) was between MURAL and POOL, with MURAL having more positive perceptions.

4.3 Discussion

Our goal was exploratory – to investigate a small number of images in order to get a sense of how sensitive performance in PassPoints is to the images used. We found that there were no striking differences in performance, either in the learning phase or the retention phase. As expected, there was a significantly higher number of incorrect password submissions in R2, and input time for incorrect and correct password submissions in R2 were longer. However, there were few significant differences among the images. There were some differences in perceptions of the image groups, with the MURAL group usually more positive.

Our sense of the results is that users can successfully use a variety of images. Nevertheless, we did observe that, although not significant, there was a trend for some images to perform worse than others. The POOL image tested most poorly in many of the analyses, whether it be learning, retention, or participant perceptions. This suggests that even if an image is usable, there are likely to be better and worse images to use as password images.

Clearly, one could find many bad images that should be avoided, for example, images with few memorable click points, such as an image with large expanses of blue sky [33] or jumbled, incomprehensible scenes [3]. Other images that one would want to avoid might be images with little color or low contrast [6]. Abstract images are also likely to be poor password images. Abstract swirls of color were used, apparently successfully, in Déjà Vu [14], but that system was based on image recognition. A swirl of color or other abstraction would probably be a poor image for a system based on clicking specific memorable area in an image. Images that are pleasant or have positive affect may support memorability [8]. Finally, images associated with the individual graphical password user may be memorable, but pose the danger that someone who knows the user would be able to guess the password.

While research from psychology helps, unfortunately our limited knowledge about the relationship of image content and memory makes choosing password images an art rather than a science. It appears that many images are probably usable and the main goal should be to avoid bad images that will confound memory. While an image with poor memory characteristics may be acceptable if frequently used, it will probably be quite susceptible to forgetting in infrequent use.

5. Conclusion

With respect to the tolerance experiment, we can conclude that the smaller tolerance of 10 x 10 pixels seriously impaired users' memory, and correspondingly increased their password input time, after one week in which the password was not used. Our interpretation of this phenomenon is that users who forgot their passwords failed in the learning phase to encode their password points in memory precisely. Generally, they were able to identify the area of their point but had not stored sufficiently precise knowledge about the points. With the small tolerance they were much less likely to click within the tolerance than users in the larger 14 x 14 pixel tolerance. This effect would be likely to decrease with long-term, regular use of the password, i.e., as their performance became more automated. However, if that precise memory decayed over a long lapse in usage, the user would again be susceptible to failure because of the small margin of error.

In the images experiment we found that there were few significant differences among several images of everyday scenes. Using guidance from psychology as well as intuition one may be able to choose images that are sufficiently good password images and avoid at least the worst images that interfere with memorability. However, further work on password images is needed to determine to what extent images have “hot spots” that attract many users to choose password points in the same small areas. If hot spots occur frequently, then they reduce entropy of the system. This phenomenon has been shown by Davis [*] in face recognition graphical passwords, but the danger may be less in our system with good choice of images to avoid hot spots. We plan to begin studying hot spots by collecting a large number of password points on multiple images.

As a final note, our results in the two studies reported here have many similarities in overall performance. In particular, we see here and in our earlier study [33] fast performance in learning, followed by fast and accurate performance in the following retention trial. This is followed in all our experiments by much poorer memory in the retention trial one week later. This pattern exists regardless of tolerance or image. In our prior study, however, we had a third retention trial 4 weeks after the second trial. In this delayed trial the memory problems decreased, i.e., there were many few incorrect password submissions. Thus, it appears that users consolidate their memory of the password over time. Perhaps difficulties they experienced in the one-week retention trial forced them to encode their passwords more precisely. We are interested in studying the process of consolidation of graphical passwords in memory more fully and in investigating the time to input graphical passwords when the user has become highly skilled. This automation did not occur in our studies because of the focus on memorability, which dictated intermittent use over time. With skilled users we would like to compare our experimental results to predictions of Fitts' Law [17].

6. ACKNOWLEDGMENTS

This work was supported in part by NSF grants CCR-0310490 and CCR-0310793.

7. REFERENCES

- [1] Adams, A. and Sasse, M.A. Users are not the enemy. *CACM* 42, 12 (1999), 41-46.
- [2] Bahrck, H.P. semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Verbal Learning and Verbal Behavior* 14 (1984), 1-24.
- [3] Biederman, I., Glass, A.L. and Stacy, E.W. Searching for objects in real world scenes. *Journal of Experimental Psychology* 97 (1973), 22-27.
- [4] Birget, J.C., Hong, d., Memon, N. Robust discretization, with application to graphical passwords. *Cryptology ePrint Archive*, <http://eprint.iacr.org/2003/168>, accessed Jan. 17, 2005.
- [5] Blonder, G.E. Graphical passwords. United States Patent 5559961, (1996).
- [6] Borges, M.A., Stepnowsky, M.A., and Holt, L.H. Recall and recognition of words and pictures by adults and children. *Bulletin of the Psychonomic Society* 9, 2 (1977), 113-114.

- [7] Boroditsky, M. Passlogix Password Schemes. <http://www.passlogix.com>. Accessed Dec. 2, 2002.
- [8] Bradley, M.M., Grenwald, M.K., Petry, M.C. and Lang, P.J. Remembering pictures: Pleasure and arousal in memory. *Journal of Experimental Psychology* 81, 2 (1992), 379-390.
- [9] Brown, A.S., Bracken, E., Zoccoli, S. and Douglas, K. Generating and remembering passwords. *Applied Cognitive Psychology* 18 (2004), 641-651.
- [10] Brostoff, S. and Sasse, M.A. Are Passfaces more usable than passwords: A field trial investigation. In *People and Computers XIV - Usability or Else: Proceedings of HCI 2000 (Bath, U.K., Sept. 8-12, 2000)*. Springer Verlag, 405-424.
- [11] Coventry, L., De Angeli, A. and Johnson, G. Usability and biometric verification at the ATM interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'03)* (Fort Lauderdale, FL, USA, April 5-10, 2003). ACM Press, New York, NY, 153-160.
- [12] Davis, D. Monroe, F. and Reiter, M.K. On user choice in graphical password schemes. In *Thirteenth Usenix Security Symposium* (San Diego, CA, USA, Aug. 9-13, 2004). <http://www.usenix.org/events/sec04/tech/davis.html>, accessed: Feb. 21, 2005.
- [13] De Angeli, A., Covernty, L., Cameron, D., Johnson, G.I. and Fischer, M. VIP: A visual approach to user authentication. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI 2002)* (Trento, Italy, May 22-24, 2002). ACM Press, New York, NY, 316-323.
- [14] Dhamija, R. and Perrig, A. Déjà Vu: User study using images for authentication. In *Ninth Usenix Security Symposium* (Denver, CO, USA, Aug. 14-17, 2000). <http://www.usenix.org/publications/library/proceedings/sec2000/dhamija.html>, accessed: Feb. 20, 2005.
- [15] Dourish, P. Security as experience and practice: Supporting everyday security. Talk given at the *DIMACS Workshop on Usable Privacy and Security Software*, July 7, 2004.
- [16] Feldmeier, D.C. and Karn, P.R. UNIX password security – ten years later. In *Advances in Cryptology – CRYPTO'89*, Lecture Notes in Computer Science 435, Springer Verlag (1990), 44-63.
- [17] Fitts, P.M. The information capacity of the human motor system in controlling amplitude of movement. *Journal of Experimental Psychology* 47 (1954), 381-391.
- [18] Hollingsworth, A. and Henderson, J.S. Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology – Human Perception and Performance* 28 (2002), 113-136.
- [19] Jain, A., Hong, L. and Pankanti, S. Biometric identification. *CACM* 43, 2 (2000), 91-98.
- [20] Klein, D. A survey of, and improvement to, password security. In *UNIX Security Workshop II Proceedings, Tenth Usenix Security Symposium* (Portland, OR, USA, Aug. 27-28, 1990), 83-86.
- [21] Mandler, J.M. and Ritchey, G.H. Long-term memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory* 3 (1977), 386-396.
- [22] Morris, R. and Thompson, K. Password security; A case study. *CACM* 22 (1979), 594-597.
- [23] Nelson, D.L., Reed, U.S., and Walling, J.R. Picture superiority effect. *Journal of Experimental Psychology: Human Learning and Memory* 3 (1977), 485-497.
- [24] Norman, D.A. *The Design of Everyday Things*. Basic Books, New York, NY, 1988.
- [25] Patrick, A.S. Long, A.C. and Flinn, S. HCI and security systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'04)* (Vienna, Austria, April 24-29, 2004). ACM Press, New York, NY, 1056-1057.
- [26] Paivio, A., Rogers, T.B. and Smythe, P.C. Why are pictures easier to recall than words? *Psychonomic Science* 11, 4 (1976), 137-138.
- [27] Real User Corporation. The Science Behind Passfaces. <http://www.realusers.com>. Accessed Dec. 2, 2002.
- [28] Rundus, D.J. Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology* 89 (1971), 63-77.
- [29] Sasse, M.A., Brostoff, S. and Weirich, D. Transforming the 'weakest link' – a human/computer interaction approach to usable and effective security. *BT Technical Journal* 19 (2001), 122-131.
- [30] Shepard, R.N. Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior* 6, 156-163.
- [31] Standing, L.P. Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology* 25, 207-222.
- [32] Weinshall, D. and Kirkpatrick, S. Passwords you'll never forget, but can't recall. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'04)* (Vienna, Austria, April 24-29, 2004). ACM Press, New York, NY, 1399-1402.
- [33] Wiedenbeck, S., Waters, J., Birget, J.C., Broditskiy, A. and Memon, N. PassPoints: Design and evaluation of a graphical password system. Submitted, (2005).
- [34] Wixted, T.J. The psychology and neuroscience of forgetting. *Annual Review of Psychology* 55 (2004), 235-269.