A Second Look at the Usability of Click-Based Graphical Passwords

Sonia Chiasson
Human-Oriented Technology Lab
& School of Computer Science
Carleton University
chiasson@scs.carleton.ca

Robert Biddle
Human-Oriented Technology Lab
Carleton University
robert_biddle@carleton.ca

P.C. van Oorschot
School of Computer Science
Carleton University
paulv@scs.carleton.ca

ABSTRACT
Click-based graphical passwords, which involve clicking a set of user-selected points, have been proposed as a usable alternative to text passwords. We conducted two user studies: an initial lab study to revisit these usability claims, explore for the first time the impact on usability of a wide-range of images, and gather information about the points selected by users; and a large-scale field study to examine how click-based graphical passwords work in practice. No such prior field studies have been reported in the literature. We found significant differences in the usability results of the two studies, providing empirical evidence that relying solely on lab studies for security interfaces can be problematic. We also present a first look at whether interference from having multiple graphical passwords affects usability and whether more memorable passwords are necessarily weaker in terms of security.

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, Experimentation.

Keywords
Usable security, graphical passwords, authentication, user study.

1. INTRODUCTION
Click-based graphical passwords, which involve clicking a set of user-selected points, have been proposed as a usable alternative to text passwords. Wiedenbeck et al. [16][17][18] conducted in-lab user studies of a proposed click-based graphical password scheme called PassPoints. While initial results were optimistic with respect to usability, they acknowledged that further work was needed to address several remaining questions. These included conducting a field study assessing the usability of PassPoints in a more realistic setting, investigating the effect of screen size on usability, examining whether hotspots cause security concerns, and looking at the effect of interference, i.e., whether having to remember multiple graphical passwords might cause memorability or usability problems.

Symposium On Usable Privacy and Security (SOUPS) 2007, July 18-20, 2007, Pittsburgh, PA, USA. (Version: May 24, 2007)

We conducted two user studies addressing these issues and re-examining earlier usability claims. Our first study was conducted in-lab to establish whether we could confirm the initial usability claims, look more closely at whether image choice had any impact, and gather click-point data. We tested 17 different images and gathered a range of 31 to 44 passwords from distinct participants on each image. Secondly, we conducted a field study where 376 students used click-based graphical passwords to access their class notes during the Fall 2006 semester. This provided realistic usage data to evaluate in terms of usability and let us examine interference, as a subset of students had two graphical passwords. For this field study, we selected two images from the set tested in the lab study.

A security analysis was conducted on both data sets, looking specifically at the emergence of hotspots, seeing whether hotspots could be predicted by automated methods, and demonstrating how collecting a small subset of passwords can be used to conduct successful dictionary attacks. This security analysis is reported separately [14]. Using these results, we subsequently evaluated an additional security issue: whether more memorable passwords (i.e., passwords for which users had a higher login success rate) were weaker from a security point of view (i.e. more easily cracked).

We also compared the results of our lab study with our field study and show that the lab study is not a reliable predictor of real-world usability. This raises significant methodological concerns in usable security since lab studies are often used as the primary means to evaluate usability.

A number of our results differ materially from previous usability studies [16][17][18]. We found that participants were remarkably accurate in entering their passwords, indicating that tolerance regions as small as 9x9 pixels may be acceptable. It also appears that the type of image impacts memorability, with some images being too difficult to use. We further found that interference appears to be a problem. Participants who had two passwords had significantly lower success rates than those who had only one.

The remainder of the paper is divided as follows. Section 2 summarizes the original PassPoints studies by others and related work on graphical passwords. Section 3 details our lab study and its results, while Section 4 describes the field study. Discussion and comparison of our two studies are presented in Section 5. Lastly, Section 6 offers our concluding remarks.

2. BACKGROUND AND RELATED WORK
Graphical passwords have been proposed as alternatives to text passwords to improve both usability and security issues. Suo et al.

[13] survey a large number of existing schemes. Psychology studies reveal that humans are better at recognizing and recalling images than text; graphical passwords are intended to capitalize on this in hopes that by addressing the memory burden on users, more secure passwords can be produced and users will not resort to unsafe practices in order to cope [8].

The original idea for click-based graphical passwords is attributed to Blonder [2] who described a scheme where users created a password by clicking on a series of predefined regions within an image. Blonder's limitation of only allowing users to select from predefined objects within the image made the password space too small to be secure. Wiedenbeck et al. [16][17][18] subsequently proposed PassPoints, an alternative scheme allowing users to click anywhere on the given image.

With PassPoints, users create a password by clicking five ordered points anywhere on the given image. To log in, users must correctly repeat the sequence of clicks, with each click falling within an acceptable tolerance of the original point. To implement this aspect, along with a scheme converting the user-entered graphical password into a cryptographic verification key, a "robust discretization" scheme was proposed [1]. It consisted of three overlapping grids (invisible to the user) used to determine whether the click-points of a login attempt were close enough to the original points to be accepted.

Wiedenbeck et al. [16][17][18] conducted three user studies examining the effects of image choice and size of the tolerance region, and comparing PassPoints-style graphical passwords with text passwords. All three studies were conducted in-lab and consisted of having users create a password and practice until they entered it correctly ten times. At the end of the session, users logged in with their newly memorized password. They returned one week later to log in again; in addition, for one study they also returned at the 6-week mark. The stated conclusion [18] was that despite the fact that graphical passwords are slower to enter than text passwords and users made more mistakes in the learning phase, both types of passwords are similarly memorable. From the second study [16], the stated conclusion was that while using a smaller tolerance square led to a larger password space, squares of 10x10 pixels were too small to be usable, with recommended tolerance regions of 14x14 pixels or larger. A third conclusion [16] was that image choice had little impact on the memorability of passwords; users performed equally well on the four images tested. The issue of "hotspots", areas on the image that users are more likely to select, were briefly considered but the suggestion was that further investigation is required to determine whether these are a problem.

Davis et al. [4] conducted a field study where students used one of two graphical password schemes, namely Face and Story, to access class material. Users selected their set of password images from among decoys. Face used only images of human faces while Story contained everyday images. One of their major conclusions was that many graphical password schemes, including Face and Story, may require "a different posture towards password selection" than text passwords, where selection by the user is the norm. Weinshall [15] reported on an in-lab user study of a proposed graphical password scheme where users identified images from their pre-determined set of secret images; but this scheme has been attacked by Golle and Wagner [6]. The attacks used a SAT solver, allowing recovery of the user's secret in a few seconds, after seeing a small number of user logins.

3. LAB STUDY

We first conducted a lab study to independently evaluate the usability of click-based graphical passwords. Our methodology differed from the original studies (see below) but still consisted of having users create and confirm a graphical password then log in using that password. We tested 17 different images with 43 participants, giving a range of 31 to 44 collected passwords on each image. The study's methodology was approved by our university's Psychology Research Ethics Committee and conducted at the university's HCI usability lab.

3.1 Methodology for the Lab Study

We used a web-based interface developed with PHP for this study. A new version of the software was developed because the PassPoints source code was not available to us. Our images were 451x331 pixels in size, the same dimensions as in the PassPoints studies. The original PassPoints studies reported using a 20x20 pixel tolerance square, however it is unclear how this was implemented since it is impossible to accurately center a 20x20 square on a given pixel. We decided on a tolerance square of 19x19 pixels centered on the original click-point. In other words, confirm and login attempts where all points were less than 10 pixels in any x- or y- direction from their corresponding original click-points were considered successful.

Since we wanted to perform analysis on the passwords collected and the exact points selected, we did not use any discretization methods [1] nor hash the passwords before storing them. We simply recorded the exact coordinates of the click-points. As in the Wiedenbeck et al. studies, we used a Windows-based desktop computer with a 19-inch screen set at a resolution of 1024x768 pixels.

In our lab study, we tested 17 different images. The images were selected to represent a variety in terms of level of detail, visual clutter, amount of colour, and content (landscapes, close-ups of objects, people, maps, etc.). Our set included the four images from the original PassPoints studies.

Participants created passwords on as many of these images as possible during their session. The number of images seen by each individual participant ranged from 9 to 17. In total, we collected a range of 31 to 44 passwords on each of the images. The maximum is greater than the total number of participants because some participants changed their password if they could not remember it. Participants were assigned a two-digit username that they used throughout the session. Wiedenbeck et al.'s interface did not require a username, but we felt that having one was more realistic.

We did not follow the exact methodology used by Wiedenbeck et al. for a few reasons. First, we did not feel that requiring users to correctly confirm their password ten times before logging in reflected a realistic usage scenario. Pilot testing revealed that this was a frustrating experience that would annoy participants. Secondly, following this procedure would not have allowed us to test multiple images due the time it took for each image.

3.1.1 Participants

Forty-three participants (25 females, 18 males) took part in this study. Data from two participants was eliminated because a

malfunctioning mouse affected their performance. This paper considers data from the 41 remaining participants. All participants were university students from various degree programs, with an even mix of graduate and undergraduate students. Ten had technical backgrounds, but none were majoring in computer security. The average age of participants was 27 years. Thirty-seven reported using the web daily while the remaining four said they were online several times a week, so all were adequately experienced with using a computer and the web. Most participants (33) indicated that they were concerned about the security of passwords or that they took steps to reduce risks, yet 37 of them admitted to reusing passwords. None had any experience with graphical passwords.

3.1.2 Task

Each participant completed a one-hour session in our usability lab. After completing the consent forms, they were introduced to the idea of graphical passwords. As part of this introduction, the experimenter showed them an image on the screen with a small superimposed square and explained that this was how accurate they needed to be with their mouse clicks when re-entering their passwords. They were advised to pretend that these passwords protected their bank information which meant that while they should pick something they could remember, they should also select passwords that would be difficult for others to guess so that no one could break into their account.

Each trial followed the steps described below. Steps 1, 2, and 5 represent the password phases on which analysis is reported later in this paper.

1. Create Phase: Participants entered their username, selected a password by clicking five consecutive points on the given image, and clicked on the Login button. Their password consisted of these five points in the specified order.
2. Confirm Phase: The same image was presented a second time and users were asked to confirm their password. They once again entered their username and password then pressed the Login button.
3. Two-questions: After successfully confirming their password, the following screen asked two 10-point Likert-scale questions: "How easy was it to create a password on this image?" and "How difficult will it be to remember your password in one week?"
4. Mental Rotations Test (MRT) puzzle: The MRT is a paper-based test used in psychology experiments as a measure of spatial ability [11]. Participants typically complete as many of the puzzles as possible within a given time (about 5 minutes). In this study, our intent was to distract participants and remove their password from working memory by clearing their "visual working memory". Psychology literature suggests that 15-30 seconds is ample time for this to occur [5]. We gave participants an MRT puzzle to solve and ensured that at least 30 seconds had elapsed before moving on. If they completed the puzzle too quickly, we gave them a second puzzle, but this happened very rarely.
5. Login Phase: Participants then logged in using their previously created password.

If participants were unable to confirm their password or log in after 2 attempts, they were allowed to change their password (in effect returning to Step 1) or if they strongly disliked the image or found it too difficult, they could skip this trial and move on to the next one. Note that contrary to the PassPoints studies, we did not display the password click-points superimposed on the image after users had selected their click-points because revealing the password on the screen seemed unrealistic in a real-world setting.

The first two trials for each participant were considered "practice" trials, with the experimenter guiding users through the process and answering any questions they may have had to ensure that users understood the tasks. Data from these two trials were discarded during analysis. Participants then completed trials with as many images as possible in the remaining time, while working at their own pace. They were allowed to take breaks as needed between trials. After approximately half an hour, the experimenter interrupted, telling them to take a break and asking them to answer a demographics questionnaire. To avoid bias on any image due to inexperience or fatigue, the order of the images was randomly shuffled so that no two participants saw them in the same order.

At the end of the session, participants completed a post-test questionnaire. This questionnaire asked about their opinion of the system and graphical passwords then asked about their strategy for selecting passwords and the types of images they preferred.

3.1.3 Data Collection

Both quantitative and qualitative data was collected during the lab study. Computer logs were generated to record each Create, Confirm, and Login attempt made by participants. Besides collecting the coordinates of the selected points, timestamps were recorded for each point as well as the total time elapsed from when the image was first displayed to when users pressed the Login button. Responses to the two questions from Step 3 were also stored.

Participants' responses to the demographics and post-test questionnaires as well as the MRT puzzles were also collected. Additionally, the experimenter sat with each participant throughout the sessions, recording any comments made by participants as they worked, any observed usability problems, and other observations. Care was taken to only ask questions such as "what did you think of this image?" in between trials so that the timings remained as accurate as possible. However if participants chose to talk during a trial, they were not discouraged.

3.2 Collected Results for the Lab Study

Only 20 out of 41 participants had time to complete all 17 images, however since the order of the images was shuffled, we obtained at least 31 created passwords for each image. In total, data from 582 trials were analyzed. In some of the results reported here, we give primary focus to the Pool and Cars images (see Figure 5 and Figure 6) since these are the images used in the second study.

Four types of statistical tests [7] for significance were used during the data analysis, each intended to determine whether the groups being analyzed were distinct from each other with respect to the factor being tested. Results from ANOVAs are reported when comparing the means across multiple groups, t-tests are used when comparing means between two groups, Mann-Whitney tests are used when comparing ordered categorical data, and Chi-square tests (χ^2) are used for non-ordered categorical data. In all cases, a value for $p < .05$ indicates that the groups being tested are

different from each other with at least 95% probability, making the result statistically significant. In the tables, a value of n.s. means that the result was "not significant"; indicating no difference between the two groups with respect to the variable being tested.

3.2.1 Success Rate

Success rates were calculated as the proportion of all attempts that were successful for a given phase. The success rates for the Confirm and Login phases are provided in Table 1. Taking all images into account, a total of 628 passwords were created. Of these, 35 passwords were created on the Pool image and 31 on the Cars image. Attempts at creating a password were all considered successful because the interface did not let users move on until they had clicked five points on the image, hence successfully creating a password.

Table 1: Success rate per phase (lab)

	Pool	Cars	All 17 Images
Confirm	33/39 (85%)	31/33 (94%)	575/748 (77%)
Login	33/33 (100%)	30/32 (94%)	560/598 (94%)

Figure 1: Success rate per phase (lab)

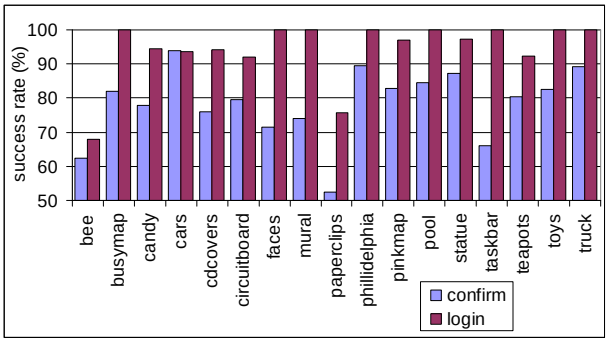


Figure 1 shows the Confirm and Login success rates for each of the 17 images. There is considerable variation between images; in fact, statistically significant differences between images are seen for both the Confirm (χ^2 (16, N=748) = 49.64, $p < .001$) and Login (χ^2 (16, N=598) = 91.44, $p < .001$) phases. For example, the Paperclips image had the worst success rate in the Confirm phase at 52% while the Cars image had a success rate of 94%. For the Login phase, the worst performer was the Bee image at 68% while several images reached success rates of 100%. This suggests that the choice of image can have substantial impact on usability, at least initially.

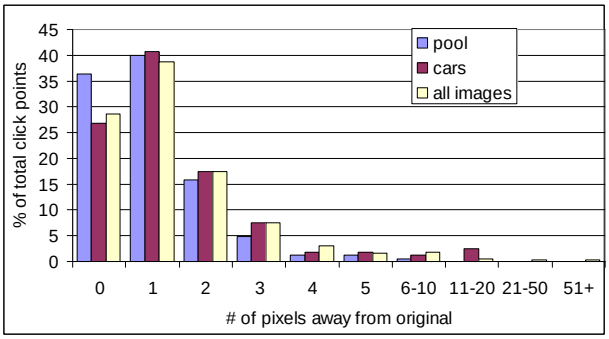
Two images had much lower success rates: the Bee and the Paperclips images. These two images were also the source of most frustration and were most frequently skipped by participants in the Confirm or Login phases. The Paperclips image consisted of a random arrangement of coloured paperclips with no obvious patterns or distinguishing features. The Bee image was a close-up photo of yellow flowers with a single bee in the center of the image. Participants disliked this image, saying that it had no obvious "clickable" points other than the bee.

From these results, we are unable to predict whether success rates for different images would converge after an initial learning curve. Success rates for the Confirm Phase are generally lower than for the Login phase. This discrepancy may be due to the fact that the Confirm phase represents the first time users re-enter their password and as such they may have forgotten their points due to inattention, may have accidentally clicked on a different point than expected, or may remember the general area (such as "the red car") but not in precise enough detail ("the left front wheel of the red car") to accurately repeat the points. From participants' comments and performance, the Confirm phase was part of the learning process; once they had successfully confirmed their password then they were more confident that they could repeat it during the Login phase. Several users stated that once they had confirmed their password successfully, then they knew it and even being distracted by the MRT did not affect their memory of it.

3.2.2 Accuracy

Participants were extremely accurate in targeting the points of their passwords. To determine accuracy, we analyzed individual click-points rather than looking at the password as a whole, therefore each password contributed 5 data points. For each point, the maximum of $|x_{original} - x_{current}|$ and $|y_{original} - y_{current}|$ was taken as the measure of accuracy. All Confirm and Login attempts were considered in the analysis, even those that were unsuccessful.

Figure 2: Accuracy for Login phase (lab)



In the Confirm phase, 96% and 94% of clicks on the Pool and Cars images respectively were within 4 pixels (1.5mm) of the original points. This means that click-points were accurate within a 9x9 pixel square. Participants were similarly accurate for the Login phase. Here, 98% of clicks were within 4 pixels for the Pool image and 94% for the Cars image. As an example, Figure 2 shows the distribution for the Login phase; the Confirm phase was very similar. There were slight variations, but participants were similarly accurate on all images. Accuracy rates appear better than success rates because success rates are based on the entire Login/Confirm attempt while accuracy rates consider individual click-points. One unsuccessful Login/Confirm attempt may have contributed four accurately entered click-points and only one incorrect click-point to the accuracy totals.

3.2.3 Times for Password Entry

As expected, it took much longer to create a password than to subsequently confirm it and log in, since participants had to initially look at the image and decide which points to select as part of their password. The total time to enter a password included typing a username (two-digits in this case), initial "think-time", clicking on five points, and clicking the Login button. Figure 3 summarizes the median total times for the Create and Confirm phases. Unfortunately, a technical glitch prevented us from gathering reliable total times for the Login phase. We report primarily median times rather than means to avoid inflated numbers due to cases where participants stopped to comment

during a trial. It also allows for comparison with our field study. The median total time for creating a password was 33 seconds (the mean time was 40 seconds), while the subsequent Confirm had a median time of 14 seconds (the mean time was 17 seconds). As shown in Figure 3, participants were quickest at creating passwords on the Truck image at 27 seconds while the Taskbar and Bee images took the longest at 42 seconds. During the Confirm phase however, times ranged only from 13 to 16 seconds.

Previous studies have found that graphical passwords take longer to enter than text passwords [13][18]. To investigate whether this extra time is due to time taken to physically move the mouse and target the click-points, we also examined the "click time", i.e., the portion of time taken from the first click-point to the last click-point. Considering all images, it took a median time of 11 seconds to click on the five points during the Create phase, and 7 seconds during Confirm and Login. Figure 4 presents the median times for each phase on each image. While these times are likely longer than typing a text password, they are probably still acceptable for entering a password.

Some images were obviously more difficult to use than others since participants took considerably longer to enter passwords on some of the images. As shown using ANOVAs, the differences in timings between images were statistically significant for all three phases (see Table 2).

Table 2: Differences between images in terms of timing (lab)

	ANOVA - Total time	ANOVA - Click time
Create	$F(15,1) = 1.94$, $p < .05$	$F(15,1) = 1.67$, $p < .05$
Confirm	$F(15,1) = 1.73$, $p < .05$	$F(15,1) = 1.66$, $p < .05$
Login	N/A	$F(15,1) = 2.46$, $p < .001$

Figure 3: Median total times per phase (lab)

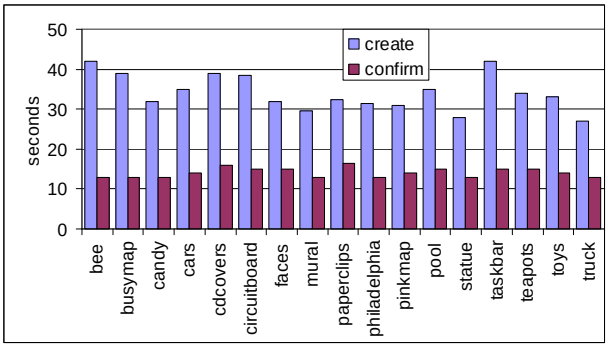
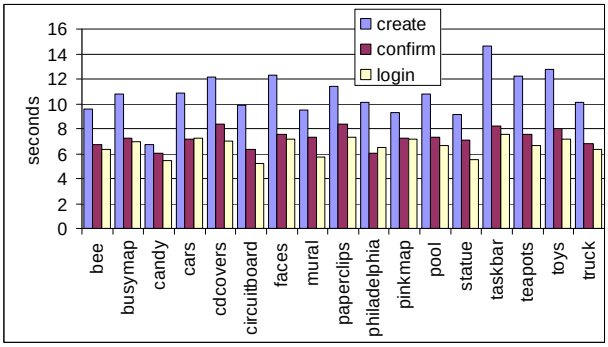


Figure 4: Median click-times per phase (lab)



3.2.4 Perceptions and Opinions

During each trial, participants answered two 10-point Likert-scale questions immediately after confirming the password. When responding to these two questions, participants rated the ease of creating the current password at 4.3/10 (the median was 4) and the ease of remembering this password after a week at 5.5/10 (the median was 6). This indicates that their immediate reaction to their graphical passwords were fairly neutral.

The post-test questionnaire contained 32 Likert-scale questions. We report only on the subset of eight questions that were also asked in the field study. For each question, a 10-point Likert scale was used with 1 indicating strong disagreement and 10 representing strong agreement with the given statement. Questions marked with an (*) used a reversed scale as a calibration to avoid bias. These scores were inverted before calculating their means and medians. In the results reported for each question, a higher value always indicates a more positive result for graphical passwords.

The eight statements were:

A. I could easily create a graphical password.

B. Someone who knows me would be better at guessing my graphical password than a stranger. (*)

C. Logging on using a graphical password was easy.

D. Graphical passwords are easy to remember.

E. I prefer text passwords to graphical passwords. (*)

F. Text passwords are more secure than graphical passwords. (*)

G. I think that other people would choose different points than me for a graphical password.

H. With practice, I could quickly enter my graphical password.

Table 3: Questionnaire responses, * = inverted (lab)

	Questions – scores out of 10							
	A	B*	C	D	E*	F*	G	H
mean	7.8	6.9	7.0	6.0	5.2	6.5	7.5	8.1
median	8	7	7	7	5	7	8	8.5

Table 3 details the means and medians for each of these questions. All of the results are in the positive range, with the exception of preference between text and graphical passwords which was neutral. Given that participants have years of experience with text passwords, it is not surprising that they did not show a stronger preference for graphical passwords. However, the positive responses to the remaining questions indicate that participants were reasonably happy with click-based graphical passwords. Since these same questions were asked in the field study (see Section 4.2.7), we were able to compare whether opinions changed with regular usage.

3.2.5 Image Preference and Click-point Selection

Participants had strong opinions of which images they liked, and especially of those they disliked. Many voiced preference for images that had "clickable points" – small, distinct areas that could easily be identified and targeted with a mouse. Structural features such as lines, repeating items, and patterns seemed to be helpful. Many people also reported using letters or numbers if they appeared on the image.

They generally disliked images that were visually cluttered or that were too similar (such as the jumbled Paperclips or the close-up

image of a uniformly coloured Circuit-board). They had trouble with the Bee image because it was mostly similar flowers and leaves with few distinct edges or distinguishing features. Most wanted to avoid clicking on the bee since it was "too obvious" but found little else that they would accurately remember.

Many reported using patterns to select their click-points, for example geometric patterns such as "four corners and the middle" or contextual patterns such as "five red cars". Some used visible angles or intersections in the image and many selected objects of distinct colours. Points with personal meaning were often selected as well; one participant commented "I have to pick something that means something to me, if I just pick something at random, it'll be much harder to remember". There was a recurring theme of needing "clickable points", although exactly what made a point clickable varied.

3.2.6 Mental Rotations Test

The Mental Rotations Test (MRT) served two purposes in this study. The first was to distract participants with a visual task that would flush their working memory and thus clear their password from memory. This strategy is often used in psychology experiments, for example by having participants count backwards for 15 to 30 seconds between tasks requiring them to remember numbers. In our case, a visual distracter was needed since graphical passwords are visual in nature.

Most participants were engrossed in completing the puzzle, often taking longer than the imposed minimum of 30 seconds. Several commented that the puzzles were harder than the passwords. It appears that the puzzles were engaging and successful at distracting participants from thinking about their password.

Secondly, since the MRT is a measure of spatial ability, its use allowed us to explore whether a higher score on the MRT can be a predictor of better performance with click-based graphical passwords. Analysis showed no correlation between the MRT results and success rates. We note that this does not necessarily indicate that spatial ability is not a factor since our test was not administered in the standard way [11].

3.3 Interpretation of Lab Study Results

Overall, the results of our lab study seem to confirm the usability of click-based graphical passwords (for further discussion, see Section 5). The success rates are high, the timings are reasonable, and participants report modestly favourable opinions of graphical passwords. With respect to accuracy, our results show that participants performed extremely well, indicating that the tolerance around the original click-points could potentially be reduced further than suggested by Wiedenbeck et al. [16] without negatively affecting usability.

Our results indicate that the choice of image had a significant impact in all areas of usability. Besides the measurable aspects, some of the more difficult images led participants to sigh and sit back on their chair, just staring at the image, obviously frustrated at trying to select points.

A direct comparison with the PassPoints results is difficult due to the differences in methodology, some of which were intentional, as discussed. However, Wiedenbeck et al. [16] conclude that image choice had little impact on usability, contrary to our results. When examining only the four images used in their study (Pool, Teapots, Philadelphia, and Mural), our lab study found no differences in success rates for the Confirm phase, but the Teapots image had 3 (92% success rate) failures in the Login phase whereas the other images had none (100% success rate). We found no differences in any of the time comparisons. So while we agree that the four images selected for Wiedenbeck et al.'s study were similar to each other, they do not appear to be a representative sample of different types of images.

Wiedenbeck et al. also concluded that a tolerance square of 10x10 pixels was too small to be usable. While we did not directly test a tolerance square of this size, our accuracy results showed that participants were extremely accurate in entering their passwords. Ninety-five percent of Login click-points were within 4 pixels of the corresponding original points and therefore would have been accepted with a tolerance square of 9x9 pixels.

It is difficult to compare times across the two studies. In one study, Wiedenbeck et al. [18] reported a time of 64 seconds to create a password, but did not report this time for their other studies. Our results showed a much shorter average password creation time of 40 seconds. This may be because our participants practiced creating two passwords beforehand. Since their Confirm phase was much different than ours, a direct comparison is not possible, and finally we do not have total times for the Login phase in our data to compare against their reported login times.

4. LONG TERM STUDY

To examine the effectiveness of click-based graphical passwords in a real-world setting, we conducted a field study where students used a graphical password to access their class notes during the Fall 2006 semester for 7 to 9 weeks. The study was reviewed and approved by our university's ethics committee for psychological research as well as the Computer Science department head and the respective instructors since it involved students from Computer Science classes.

4.1 Methodology for the Field Study

A web-based PassPoints-style graphical password system was built where students logged in to access the instructor's class notes. The system was available from mid-October to mid-December, with students logging on whenever they wanted to access their class material. Students who preferred not use a graphical password could opt-out and create a text password instead. In total, 376 students created graphical passwords and 25 created text passwords.

Students were introduced to the system through a combination of demonstrations during class time and tutorials, email instructions, and FAQ/Help on the system's web page. We received only a handful of requests for technical support throughout the study.

The first time students accessed the system, they entered secondary identification information, created a secret question in case they needed to change their password, and proceeded to create and confirm their click-based graphical password on an assigned image (see Section 4.1.2). A small square directly above the image reminded them of the accepted tolerance for their points. Passwords consisted of an ordered series of five unique points as in our lab study.

4.1.1 Participants

Students from three first-year undergraduate Computer Science (CS) classes were invited to participate in this study. One class was for students who were not CS majors while the other two

were primarily for students intending to major in CS. We received consent from 191 unique students to use their data in our study (124 CS students and 65 non-CS students). Of these, 37 students were in two of the classes and had two different accounts (with different images). Therefore we have data from 228 different accounts. These 228 accounts will be used for all further analysis.

4.1.2 Study Design

A two-dimensional between-participants design was used. Participants were randomly assigned to different experimental conditions with no consideration given to which class they were enrolled, except in the cases where participants were in two classes. Both the image and the required accuracy were varied. One group was given a tolerance square of 13x13 pixels and the other a tolerance of 19x19. The 19x19 square was consistent with our lab study. Students who were in both CS classes were assigned a different image for each class but the size of their tolerance square was kept consistent.

Only two images were selected from our earlier lab study: the Pool and Cars images (Figures 5 and 6). These images had reasonable usability results and differed in their number of hotspots based on a security analysis reported separately [14]. The Pool image contained several large hotspots while the Cars image did not. The Pool image was also selected because we wanted to test one of the original PassPoints images.

The number of participants per group is given in Table 4. The size of the experimental groups are uneven because participants were assigned to groups at the beginning of the study, before we knew who would give consent to use their data.

Figure 5: The Cars image [3]

Figure 6: The Pool image [10]

The two images were the same size as in previous studies, namely 451x331 pixels. However, since students were allowed to log in from anywhere with web access, we could not control for screen size or resolution. We were nonetheless able to record their screen resolution each time they entered a password as this information is retrievable from the browser.

Table 4: Number of students per experimental condition (field)

	13x13 Tolerance	19x19 Tolerance
Pool image	63	53
Cars image	61	51

4.1.3 Data Collection

As with the lab study, we logged each password creation, confirm, and login attempt, including click-point coordinates, timestamps, as well as screen resolution. We stored the exact pixel coordinates of each point.

At the end of the semester, we asked students to complete an online questionnaire. The questionnaire included demographic questions and questions about their perception and opinion of click-based graphical passwords. Students who had two accounts were allowed to answer the questionnaire once for each account, since they may have had different responses for each image. We received 109 responses; 94 of these were from unique students.

4.2 Collected Results for Field Study

Table 5 summarizes the usage data for the field study. Participants attempted to login an average of 18 times throughout the semester and created an average of 2.6 passwords (i.e., changed their password 1.6 times). Usage was relatively consistent throughout the entire semester. The student who attempted to login most frequently did so 65 times. It should be noted that these numbers take into account all attempts, including those that were unsuccessful.

Table 5: Attempts per participant for each phase (field)

	Create	Confirm	Login
Mean	2.6	3.6	18
Median	2	2	15
Maximum	11	17	65

4.2.1 Success Rate

Success rates were calculated as the number of successful attempts across all attempts for a given phase. We decided that this was a more representative measure than calculating success rates on a per participant basis since a participant who logged in only once throughout the term could have a success rate of 100% which is rather misleading. Overall success rates for both images are provided in Table 6. The difference between images was not statistically significant during the Confirm phase, but was significant during Login (χ^2 (1, N=3443) = 16.42, $p < .001$), perhaps indicating that the choice of image does affect the memorability of passwords over time. Here, the success rates seemed to indicate that Cars was more memorable than Pool.

Participants were allowed to change their passwords at any point, provided that they entered their secondary identification information and answered their preset secret question. For the purpose of our analysis, change password attempts are treated the same as original Create attempts since the result in both cases is a new password. Once again, an attempt to create a password was only accepted once five click-points were selected, so 100% of attempts to create a password were considered successful. In total, 265 passwords were created for Pool and 216 for Cars. Of these, 149 (56%) were a result of changing a password on the Pool image in comparison to 104 (48%) for the Cars image. On the Pool image, 49% of participants created only one password and

18% created four or more. Of those using the Cars image, 43% kept the same password all term while 11% created four or more passwords.

Table 6: Success rate per phase (field)

	Pool image # success / total attempts	Cars image # success / total attempts
Confirm	207 / 388 (53%)	170 / 293 (58%)
Login	1461 / 1880 (78%)	1301 / 1563 (83%)

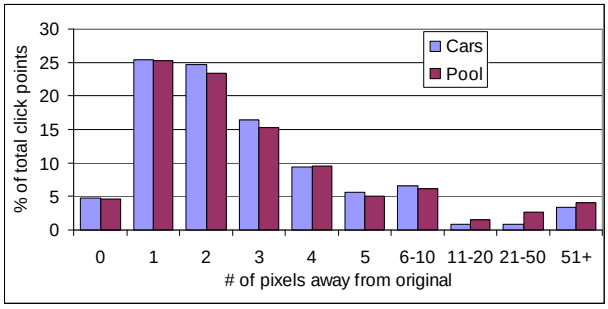
Success rates were considerably lower than in the lab study. Upon closer examination of the Login attempts, we found that success rates did improve with practice, although never reaching the levels attained in the lab study. For example, the initial success rate across all students was 76%, rising to 88% when considering only login attempts beyond the 30th attempt for students who logged in at least 30 times.

4.2.2 Accuracy

Participants were once again remarkably accurate in entering their passwords. As with the lab study, we analyzed click-points individually rather than looking at whole passwords.

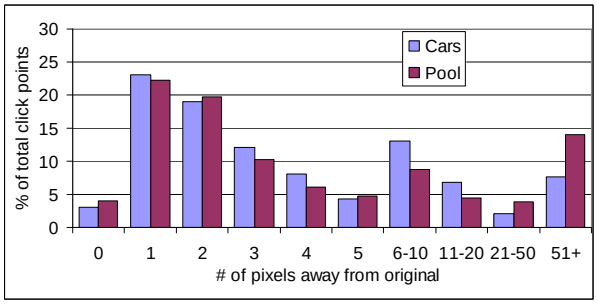
As shown in Figure 7, 78% of clicks on the Pool image for the Login phase were within 4 pixels (1.5mm) of the original point (i.e., within a 9x9 pixel tolerance square), while 80% of clicks on the Cars image fell within 4 pixels. Assuming that clicks further than 50 pixels away were forgotten points, only 4% and 3% of points were forgotten for the Pool and Cars images respectively.

Figure 7: Accuracy for Login phase (field)



Looking at Figure 8, it is apparent that confirming the password is part of the learning process as participants were considerably less accurate in entering their passwords. For the Confirm phase, 62% and 65% were within 4 pixels (i.e., within a 9x9 pixel tolerance square) for the Pools and Cars images respectively. People were also more likely than in the Login phase to forget their points altogether: 14% of points were forgotten on the Pool image and 8% of points were forgotten on the Cars image.

Figure 8: Accuracy for Confirm phase (field)



There is no statistically significant difference in terms of accuracy between the two images for the Confirm phase, but there is for the Login phase ($Z = 3.01$, $p < .01$).

4.2.3 Effect of Size of Tolerance Square

Since participants were so accurate in entering their passwords, the size of the tolerance square had little impact on success rates. For the Pool image, having different sized tolerance square had no impact on the success rates for either the Confirm or Login phases (see Table 7). The Cars image similarly showed no difference for the Confirm phase, but for the Login phase participants were significantly more likely to succeed with the larger 19x19-pixel square; however both tolerances still had success rates of above 80%, indicating little practical difference.

Table 7: Effect of size of tolerance square on success rate (field)

		13x13 Tolerance	19x19 Tolerance	χ^2
Confirm	Pool	126/245 (51%)	81/143 (57%)	n.s.
	Cars	95/170 (56%)	75/123 (61%)	n.s.
Login	Pool	790/1018 (78%)	671/862 (78%)	n.s.
	Cars	640/790 (81%)	661/773 (85%)	χ^2 (1, N=1583)=5.67, p<.05

Interestingly, participants were more accurate in entering their click-points during the Login phase when they had a smaller tolerance square. Telling them that they needed to be accurate actually improved their accuracy in the field while having little impact on their success rates. As accuracy distributions are similar to those reported in section 3.2.2, only the number of click-points within 4 pixels is reported in Table 8 although the Mann-Whitney tests take the entire data set into account.

Table 8: Effect of size of tolerance square on accuracy (field)

		13x13 Tolerance: <= 4 pixels	19x19 Tolerance: <= 4 pixels	Mann-Whitney
Confirm	Pool	781/1225 (64%)	431/714 (60%)	n.s.
	Cars	549/850 (65%)	405/615 (66%)	n.s.
Login	Pool	4174/5090 (81%)	3164/4305 (73%)	Z = 13.60, p < .001
	Cars	3289/3950 (84%)	3008/3860 (78%)	Z = 5.14, p < .001

To further examine whether the size of the tolerance square had an effect on performance, we looked at the click-time from the first to last point. If those who had a smaller tolerance square were actively trying to be more careful in targeting, we would expect to see increased click-times. However we found no statistical differences in the click-times between the two tolerance groups for either image, further indicating that participants' performance was not impacted by having a smaller tolerance square.

4.2.4 Effect of Screen Resolution

We could not control for physical screen size or screen resolution since participants could use the system from any web-enabled computer. This reflected a realistic usage scenario for most password systems. We were able to record screen resolution and examine whether it had any effect on user performance. The resolutions ranged from 256 000 pixels to 2 304 000 pixels. We divided them into two groups: one million pixels or less and greater than one million pixels.

Table 9: Effect of screen resolution on success rate (field)

		Low Res. (<= 1 million)	High Res. (> 1 million)	Mann-Whitney
Confirm	Pool	109/216 (50%)	98/172 (57%)	n.s.
	Cars	81/161 (50%)	89/132 (67%)	Z = 2.95, p < .05
Login	Pool	1000/1268 (79%)	460/611 (75%)	n.s.
	Cars	725/875 (83%)	575/687 (84%)	n.s.

Screen resolution had no impact on success rates for the Login phase. There is a significant difference for the Confirm phase of the Cars image, with a higher success rate for the higher screen resolution. It is not clear why this occurred. Since we could not record physical screen size, these results do not provide clear evidence that users perform equally well regardless of screen dimensions, however it does suggest that click-based graphical passwords are usable within the typical range screen resolutions.

4.2.5 Times for Password Entry

Participants were able to create their passwords relatively quickly, with a median total time of 25 seconds for Cars and 30 seconds for Pool. Total times for the Confirm and Login phases were surprisingly consistent, with median times varying between 13 and 15 seconds across both phases. Figure 9 presents the total times for each phase of the Cars and Pool images using Box-and-Whisker graphs. The boxes indicate the Inter-Quartile Range (IQR – the interval between the 25th and 75th percentiles) while the whiskers show the overall range of data. The thick line within the boxes indicates the median time for each phase and outliers are shown as empty circles.

Figure 9: Median total times per phase (field)

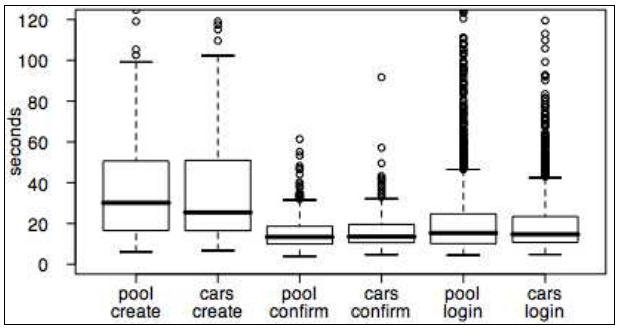
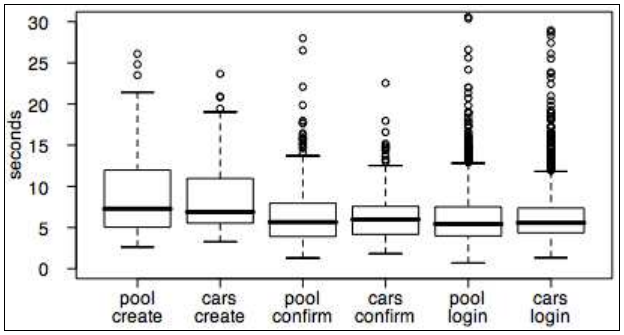


Figure 10: Median click-time per phase (field)



Mean times did not provide an accurate snapshot of the data in this case due to outliers with very high times. For example, a few Login attempts were measured in days rather than seconds. Since participants were not using the system in a controlled setting, they may have opened the login page, turned their attention elsewhere, and later returned to continue logging in. For this reason, median times are more representative.

As shown in Figure 10, participants were very quick in actually targeting and entering their click-points. When considering only the click-time from the first to the fifth point, the Create phase took slightly longer than the other phases (7 seconds) while the median click times for the Confirm and Login phases were between 5 and 6 seconds.

4.2.6 Interference

Having multiple passwords affected performance. Students who had two passwords had higher success rates in the Confirm phase (statistically significant for the Pool image; see Table 10). It appears that the extra practice at creating and confirming a password improved their performance.

During the Login phase however, interference negatively affected success rates. Students were much more likely to log in correctly when they only had one password to remember. As shown in Table 10, the difference in success rates due to the presence or lack of interference is statistically significant for both images. For example, those who only had a password on the Cars image had a success rate of 87% but those who had two passwords had a success rate of only 71% for the Cars image. This indicates that having to remember two unique passwords on different images negatively affects long-term memorability; this finding is troublesome if graphical passwords were to become widely used.

Table 10: Effect of interference on success rate[1] (field)

		No Interference	Interference	χ^2
Confirm	Pool	139/284 (49%)	63/99 (64%)	χ^2 (1, N=383)=6.36, p<.05
	Cars	108/193 (56%)	62/100 (62%)	n.s.
Login	Pool	1224/1541 (79%)	226/319 (71%)	χ^2 (1, N=1860)=11.33, p<.001
	Cars	1053/1216 (87%)	248/347 (71%)	χ^2 (1, N=1563)=44.26, p<.001

We examined more closely the data from the interference group. Specifically, we looked at the initial password created on each image to see whether their ability to confirm their password improved for the second image since they had already practiced the process with the first image. Looking only at the initial password created on each image, we uncovered that students had higher success rates for the Confirm phase for their second image (67% success rate) than on their first image (60% success rate). However, the difference did not reach statistical significance.

4.2.7 Perceptions and Opinions

At the end of the semester, participants were asked to complete an online questionnaire and 109 participants responded. We report primarily on the questions that correspond with the lab study.

Participants' opinions were neutral or mildly favourable for click-based graphical passwords in all questions except for their preference between text and graphical passwords. For this question, they strongly preferred text passwords. A summary is provided in Table 11 (see section 3.2.4 for list of questions).

[1] One participant had the same image for both classes. His data is excluded from our analysis of interference.

Table 11: Questionnaire responses, * = inverted (field)

	Questions, score out of 10							
	A	B*	C	D	E*	F*	G	H[2]
mean	6.1	6.9	6.1	5.5	3.2	5.2	6.6	6.4
median	6	7.5	7	5	3	6	7	7

In the written responses, the most common concern was shoulder-surfing. On the other hand, many felt that it would be difficult for someone to simply guess their points. Several mentioned difficulty remembering their points because they did not log in frequently enough while others said that once they had learned their password, it was easy to remember. As with the lab study, they reported selecting "easy-to-remember" points, geometric patterns (such as items in a line or a circle), colour patterns, and personally meaningful items.

4.2.8 Usability versus Security

A companion paper [14] carried out a security analysis of hotspots within the images and examined whether passwords created by a small subset of users can be leveraged to generate a successful attack against other users. Collected passwords from 35 users (Pool image) and 33 users (Cars image) in the lab study were used to determine hotspots, from which a dictionary of candidate passwords was generated. The dictionary entries were then compared to the final user-created passwords in the field study (i.e., if users changed their passwords during the semester, only the latest password was examined); any passwords where users failed to log in at least once were further eliminated. The rationale for examining this subset was that final passwords may be more indicative of what people would eventually select as memorable passwords.

The results are worrisome from a security viewpoint: the attack [14] correctly guessed 41/112 (Pool image) and 22/109 (Cars image) passwords. Taking into account all login attempts for the tested passwords, we see a statistically significant difference in the success rates between those passwords that were cracked and those that were not ($\chi 2$ (1, N=2781) = 4.67, $p < .05$). Contrary to our expectations however, the cracked passwords actually had a lower login success rate (84%) than those that were not cracked (88%).

If success rate is taken as a measure of memorability, our small sample indicates that more memorable passwords were not any easier to guess[3] than less memorable passwords. However, a larger sample, for example looking at all created passwords rather than just the final passwords, may reveal different results.

4.3 Interpretation of Field Study Results

Most people chose to use their graphical passwords throughout the semester rather than opting-out and selecting a text password, something we found encouraging in terms of usability. However the lower success rates and accuracy scores are a cause for concern, especially when combined with the fact that most people reported a preference for text passwords. Overall our interpretation was that the graphical passwords were reasonably usable, serving as well as the more familiar text passwords, but these results are still much less positive than those of the original PassPoints studies.

The effect of interference is also cause for concern since it is likely that in a real-world setting, people would have more than one password. Independent of interference, it is likely that users would resort to coping strategies that would further weaken security as they do with text passwords. In fact, many reported that they would be more likely to use geometric patterns to try and have similar passwords on each image. Interference is discussed by Wiedenbeck et al [18] and by Monrose and Reiter [8] as a potential concern but our study provides the first empirical evidence that interference is in fact a problem.

The security of graphical passwords is also questionable; we expect that the passwords would be even more vulnerable to more advanced pattern-based attacks than found to date [14] since so many users reported using colour and geometric patterns. We hypothesize that the passwords guessed in such an attack would correlate with those passwords that have higher success rates and that are more memorable.

5. FURTHER DISCUSSION

Click-based graphical passwords offer an interesting alternative to text passwords, but their usability and security must be assessed before they can be deployed as authentication mechanisms. In this section, we compare the results of our two studies and discuss how methodological factors may have influenced our results.

5.1 Comparison of Lab and Field Studies

The usability results of our two studies revealed interesting differences. The lab study provided much more positive results than the field study, calling into question the validity of only conducting lab studies for security interfaces and highlighting the risk of overstating the usability of these interfaces.

As shown in Table 12, there were statistically significant differences in the success rates and accuracy results between the lab and field studies, with the field study resulting in less positive results in both cases. This indicates that the lab study is not a good predictor of these usability aspects. With respect to password-entry times however, the field study had similar or shorter times than the lab study. For example, click-times were shorter for the Login phase in the field study (mean of 5.5 seconds) than in the lab study (mean of 7 seconds), a result that is statistically significant ($t(3403)=2.02$, $p<.05$).

There are a few possible reasons for the discrepancies between the lab and field studies. The lab study gave participants more concentrated practice with creating and confirming passwords since they performed these tasks several times within an hour. Our lab participants also had two "practice" trials where they could ask questions and become accustomed to the process before starting the real trials. In contrast, participants in the field study received an explanation and instructions, but did not have a chance to rehearse on practice images before attempting to create and confirm their real password. We felt that requiring participants to create "practice" passwords before creating their own was impractical in a realistic setting and this may partially account for the discrepancies in success rates when compared to the lab study. However our analysis also showed that while

[2] We recorded only 17 responses in the field study for question H due to a programming error.

[3] We note that this result is specific to the attack methods used for cracking, and better or other attack methods may yield different results.

success rates did improve with practice in the field study, they still did not reach the levels observed in the lab study.

Table 12: Differences in success rate and accuracy: lab vs. field

		Success Rate χ^2	Accuracy Mann-Whitney
Confirm	Pool	χ^2 (1, N=427)=14.07, p<.001	Z = 13.81, p < .001
	Cars	χ^2 (1, N=326)=16.19, p<.001	Z = 10.74, p < .001
Login	Pool	χ^2 (1, N=1913)=9.42, p<.001	Z = 13.64, p < .001
	Cars	n.s.	Z = 10.47, p < .001

Secondly, the Login phase for the lab study occurred shortly after the password was created and confirmed. Although we attempted to distract participants with MRT puzzles, the immediacy of logins likely contributed to the high success rates. As logins for the field study spanned across several weeks; participants had ample time to forget their password between login attempts.

Finally, passwords were the focus of the lab study. Participants were actively engaged in the process and it was their primary task. In the field study, the primary task was accessing class notes, while logging on was a secondary. This shift to a secondary task likely affected the amount of attention paid to the task and the importance accorded to getting it correct, even though errors hindered progress towards the primary task. This also likely partially accounts for the faster click-times as participants were trying to quickly move on to accessing their class notes.

Table 13: Differences in perception: lab vs. field

Question	Mann-Whitney
A. Easy to create	Z = 2.99, p < .01
B. Guessing	n.s.
C. Easy to log in	n.s.
D. Easy to remember	n.s.
E. Preference	Z = 4.29, p <.001
F. Strength	n.s.
G. Uniqueness	Z = 2.23, p < .05
H. Speed	Z = 2.79, p < .01

Differences were also apparent in users' opinions. Out of the eight Likert-scale questions, statistically significant differences were observed in half of responses (see Table 13). Specifically, participants of the lab study felt more positive about how easy it is to create graphical passwords, about the uniqueness of their password, about the speed of entering their password with practice, and finally they rated graphical passwords more favourably than text passwords. These differences likely reflect a changing opinion over time. Initially graphical passwords were a novel and interesting idea, however participants in the field study had a chance to incorporate these passwords into their real life and as such gain a better grasp of the strengths and limitations of click-based graphical passwords.

5.2 Characteristics of the Studies

As with any research study, methodological decisions affect the results. To put the results in context, we discuss the limitations of our studies and identify the actions taken to compensate.

5.2.1 Lab study

Our short-term study was conducted in a lab environment where participants were knowingly and actively participating in a research study. Creating passwords was their primary task and they repeated this task multiple times while being observed by the experimenter. This is necessarily an artificial environment but one that can still provide valuable insight into usability.

Our participants knew that the purpose of the study was to evaluate the usability of a "new" kind of password. In an effort to help us, they may have been extra diligent. While this may overestimate the success rates, it still shows that participants could successfully use graphical passwords if they applied themselves. On the other hand, if they were trying too hard to perform perfectly, we would expect a slowdown in the timings as they carefully tried to target each point. We saw no such delay. We further tried to mitigate problems by shuffling the order of the images to avoid bias, adding MRT puzzles as distracters, and incorporating breaks.

A common limitation of such research studies is that participants are taken from a relatively narrow population pool, often consisting of first year undergraduate students. Our participants were indeed students but they were from a broad range of fields and half were graduate students, raising the average age to 27 years. Our only pre-screening criterion was that participants should not be majoring, or experts, in computer security.

Despite the limitations of such lab studies, we believe that our methodology was sound and that it provided us with useful results to guide the design of the field study, revisit previous lab usability claims, and provide data for a diverse set of images.

5.2.2 Field study

The goal of our field study was to gain an understanding of how well click-based graphical passwords work in a large-scale, practical, real-world scenario. It certainly provided a more realistic view of the usability of this method of authentication.

Participants were using graphical passwords as part of their regular activities. They knew that we would be analysing this data for our research, but they were not being observed in-person and were not in a laboratory setting which should lead to more natural behaviour on their part.

There is further evidence that participants were behaving naturally and not simply feeling obligated to participate. While the course instructors endorsed the study, they did not push students to participate and were not tied to the study. Approximately two-thirds of students gave us consent to use their data, exceeding our expectations since it required that students send email through the official university system which most students ignore. Those who were uncomfortable with participating simply did not consent. And finally, questionnaire responses were not overly positive, so they were likely being truthful.

The population sample in the field study was less diverse than in the lab study. Two-thirds of students were in a first-year undergraduate course for those majoring CS. Of those who answered the questionnaire, the mean age was 22. If anything, this group is probably more willing to accept new technology than other populations.

One potential criticism of this study is that while the passwords were protecting something valuable to participants, they were not protecting something private or personal. This was an intentional design decision since to-date there had been no in-depth security analysis conducted on click-based graphical passwords. We did not want to risk exposing private or personal information and as a result, participants may have selected weaker passwords. On the

other hand, people do not necessarily select strong passwords for high-value accounts either because they are unaware of the risks, security is a secondary task, or they do not understand what makes a secure password. Schechter et al. [12] present a study of security indicators for online banking where some participants used their real banking credentials, however these credentials were never recorded and the study was conducted in a lab environment. Conversely, our field study was not in a controlled environment and we were specifically examining the passwords so their approach would not have been appropriate.

6. CONCLUSION

We present the results of two usability studies of click-based graphical passwords. The initial lab study revealed mostly positive results and led to a larger field study to see how click-based graphical passwords worked in practice.

The lab study confirmed earlier work that the usability of these passwords was good in terms of success rates and password-entry times and that participants' opinions were favourable. We additionally showed that participants were more accurate in targeting their click-points than previously suggested; indicating that smaller tolerance squares may be acceptable. Finally, contrary to previous work, we found that the choice of image significantly influenced success rates.

The field study represented the first large-scale, real-world study of click-based graphical passwords, showing that graphical passwords were adequate in terms of usability for real tasks. Password entry times were acceptable, accuracy was not quite as high as in-lab but still very good, and success rates improved with practice although they never reached those seen in lab. Participants' opinions of graphical passwords were lower than in the lab study, suggesting that opinions worsened with real-world usage. We found several legitimate concerns with adopting graphical passwords as a means of authentication. We provided the first empirical evidence that interference from having to remember multiple graphical passwords is problematic. Participants also reported using patterns in selecting their passwords, suggesting increased susceptibility to guessing attacks.

The differences between the lab and field studies also raise methodological concerns in usable security. So far, lab studies are the most common form of usability evaluation and while others have cautioned that these were inadequate in providing realistic usability data, our two studies provide empirical evidence of this problem.

7. ACKNOWLEDGMENTS

We thank Julie Thorpe for her collaboration in this project, the participants of our lab study, and Prosenjit Bose, Louis D. Nel, Weixuan Li and their students for participating in our field study. The first and second authors are supported in part by the "Legal and Policy Approaches to Identity Theft" project funded by the Ontario Research Network for E-Commerce. The third author acknowledges NSERC for funding an NSERC Discovery Grant and his Canada Research Chair in Network and Software Security.

REFERENCES

[1] Birget, J.C., Hong, D., and Memon, N. *Graphical Passwords Based on Robust Discretization*. IEEE Transactions on Information Forensics and Security 1(3), pp. 395-399, 2006.

[2] Blonder, G.E. *Graphical Passwords*. United States Patent 5,559,961, 1996.

[3] Britton, Ian. http://www.freefoto.com Accessed Feb. 2007.

[4] Davis, D., Monrose, F., and Reiter, M.K. *On User Choice in Graphical Password Schemes.* USENIX Security 2004.

[5] Goldstein, E.B. *Cognitive Psychology*. Wadsworth Publishing, pp. 150-161, 2006.

[6] Golle, P. and Wagner, D. *Cryptanalysis of a cognitive authentication scheme (extended abstract)*. 2007 IEEE Symposium on Security and Privacy.

[7] Heiman, G.W. *Basic Statistics for the Behavioral Sciences*. Houghton Mifflin Company: Boston, MA. 1992.

[8] Monrose, F. and Reiter, M.K. *Graphical Passwords*. Chapter 9 in Security and Usability: Designing Secure Systems That People Can Use. L.F. Cranor and S. Garfinkel (eds). O'Reilley, 2005.

[9] Nelson, D.L., Reed, U.S., and Walling, J.R. *Picture Superiority Effect*. Journal of Experimental Psychology: Human Learning and Memory 3, pp. 485-497, 1977.

[10] PD Photo. http://pdphoto.org Accessed Feb. 2007.

[11] Peters, M. *Revised Vandenberg & Kuse Mental Rotations Tests: forms MRT-A to MRT-D*. Technical Report, Department of Psychology, University of Guelph, 1995.

[12] Schechter, S.E., et al. *The Emperor's New Security Indicators: An evaluation of website authentication and the effect of role playing on usability studies.* 2007 IEEE Symposium on Security and Privacy.

[13] Suo, X, Zhu, Y., and Owen, G.S. *Graphical Passwords: A Survey*. ACSAC 2005.

[14] Thorpe, J. and van Oorschot, P.C. *Human-Seeded Attacks and Exploiting Hot-Spots in Graphical Passwords*. USENIX Security 2007 (to appear).

[15] Weinshall, D. *Cognitive Authentication Schemes Safe Against Spyware (short paper).* 2006 IEEE Symposium on Security and Privacy.

[16] Wiedenbeck, S., Waters, J., Birget, J.C., Brodskiy, A., and Memon, N. *Authentication Using Graphical Passwords: Effects of Tolerance and Image Choice*. SOUPS 2005.

[17] Wiedenbeck, S. et al. *Authentication Using Graphical Passwords: Basic Results.* HCII 2005.

[18] Wiedenbeck, S. et al. *PassPoints: Design and longitudinal evaluation of a graphical password system*. International Journal of Human-Computer Studies 63, pp. 102-127, 2005.