Exploring Usability Effects of Increasing Security in Click-based Graphical Passwords

Elizabeth Stobert[1], Alain Forget[2], Sonia Chiasson[2],
Paul van Oorschot[2], Robert Biddle[2]
[1]Department of Psychology, [2]School of Computer Science
Carleton University, Ottawa, Canada
estobert@connect.carleton.ca, aforget@scs.carleton.ca, chiasson@scs.carleton.ca,
paulv@scs.carleton.ca, robert_biddle@carleton.ca

ABSTRACT

Graphical passwords have been proposed to address known problems with traditional text passwords. For example, memorable user-chosen text passwords are predictable, but random system-assigned passwords are difficult to remember. We explore the usability effects of modifying system parameters to increase the security of a cued-recall, click-based graphical password system. In this system, users create passwords consisting of one click per image on a click-dependent sequence of images. Generally, usability tests for graphical passwords have used configurations resulting in password spaces smaller than that of common text passwords. Our two-part lab study compares the effects of varying the number of click-points and the image size, including when configurations provide comparable password spaces. We use the J-statistic for comparison of relative clustering, which is known to impact security. For equivalent spaces, no usability advantage was evident between more click-points, or a larger image. This is contrary to our expectation that larger image size (with fewer click-points) might offer usability advantages over more click-points (with correspondingly smaller images). The results suggest promising opportunities for better matching graphical password system configuration to device constraints, or capabilities of individual users without degrading usability. For example, using more click-points on smart-phone displays (where larger image sizes are not possible).

Categories and Subject Descriptors

K.6.5 [**Computing Milieux**]: Security and Protection—*Authentication*

General Terms

Experimentation, Human Factors, Security

Keywords

Authentication, Graphical Passwords, Usable Security

Technical Report TR-10-06, March 8, 2010
School of Computer Science, Carleton University, Ottawa, Canada
`http://www.scs.carleton.ca/research/tech_reports`
.

1. INTRODUCTION

The problems of knowledge based authentication, typically text-based passwords, are well known. Users often create memorable passwords that are easy for attackers to guess, but strong passwords assigned by the system are difficult for users to remember [29]. Users also tend to reuse passwords across many accounts [18] and this increases the potential impact if one account is compromised. Alternative solutions to text passwords seek to improve memorability without reducing security. Graphical passwords [4, 30] use images instead of text for authentication. They attempt to leverage the *pictorial superiority effect* [26] which suggests that humans are better able to remember images than text. Some graphical password systems also provide *cueing* [9], whereby a memory retrieval cue is provided to help users remember and distinguish their passwords. In this paper we explore methods for increasing the security of a cued-recall graphical password scheme[1].

We chose to study Persuasive Cued Click-Points (PCCP), a click-based graphical password system in which users select click-points on more than one image [4, 35, 21]. Each image shown provides a cue to help the user remember the corresponding click-point. PCCP has been shown to have good usability [5], while avoiding hotspots and patterns which can affect the predictability, and hence the security, of other click-based graphical password systems [6].

The security threat that we address involves attackers attempting to guess passwords. This danger arises when the total number of possible passwords is small, or when attackers can predict likely passwords. The design of PCCP reduces the predictability of passwords by influencing users during password creation. The number of possible passwords with its standard configuration is 2^{43}, slightly less than that of 7-character random text passwords. A gap in previous literature is that usability tests for graphical password schemes (in general) have only been carried out for configurations with password spaces smaller than that of common text passwords. To fill this gap, we explored increasing security in PCCP, conducting a study modifying two parameters: the size of the images presented, and the number of click-points in each password. The study included 83 participants who completed two sessions scheduled two weeks apart. Our results show that both manipulations affect the

[1]An early version of part of this paper, based on preliminary data, appears as an extended abstract in the ACM CHI 2010 student research competition.

usability of the system and memorability of the passwords. Moreover, when adjusted to provide the same level of security, both manipulations have similar effects on usability and memorability. This suggests that when increasing security, constraints of devices and user preferences might be accommodated. For example, when designing for mobile devices, smaller images and more click-points would be used because of the constraints of the device.

The remainder of this paper is organized as follows: we first provide some background on graphical passwords in general, and more detail on PCCP. We then introduce our study methodology, and its results. Finally, we discuss the implications of the results and offer our conclusions.

2. BACKGROUND

Graphical password systems [4, 30] are a type of knowledge-based authentication that rely on the human ability to better recognize and remember images than textual or verbal information [26]. They fall into three main categories:

Recall: (also known as drawmetric [11]) Users must recall and reproduce a secret drawing on a blank canvas (which may include grid-lines for guidance). Example systems include Draw-A-Secret [22] and Pass-Go [31].

Recognition: (also known as cognometric [11] or search-metric [27]) Users must recognize and identify images belonging to their previously memorized portfolio from within a larger set of decoy images. Example systems include PassFaces [10] and Déjà Vu [13].

Cued-recall: (also known as locimetric [11]) Users must identify and target previously selected locations within one or more images. The images act as memory cues to help recall these locations. Example systems include PassPoints [35] and Persuasive Cued Click-Points [5].

Other approaches to authentication are token-based systems and biometrics. These also have potential drawbacks, such as risks of loss, and privacy implications [23]. Another approach to knowledge-based authentication is the use of password managers where usability and the dangers of centralization remain challenges [8].

In this paper, we focus on a cued-recall, click-based system, Persuasive Cued Click-Points [5]. In PCCP, a user is presented with a number of images in sequence, and must choose one click-point per image. The first image is assigned by the system, but each subsequent image is determined by the user's previous click. In other words, clicking on different locations on an image results in different next images. This provides users with feedback about the correctness of their password entry attempt — if they see the correct next image, they can be fairly certain they have selected the correct click-point. However, this feedback is not useful to attackers because they do not know the correct sequence of images. As with other click-based graphical passwords [4, 35], the user is not expected to repeat exact pixel selections. Thus, an invisible *tolerance square* is defined around each click-point so that any of the enclosed pixels are considered acceptable. In a variation, the grid is visible to users [3]. Clicking within the correct tolerance square displays the correct next image.

Earlier click-based password schemes have a security weakness which makes passwords easier for attackers to predict. Users tend to select similar locations on images, forming *hotspots* [20, 15, 32, 28]. They also tend to select their click-points in predictable geometric patterns [6, 28]. To help create more secure passwords, PCCP includes "persuasive" elements. As shown in Figure 1, the system assists users only during password creation by providing a *viewport* that highlights a random part of the image. Users must select a click-point within this viewport. If users are unable to find a memorable point in the current viewport, they may press the *shuffle* button, which randomly repositions the viewport. Studies [5, 6] show that this random viewport, together with the shuffle button, causes click-points to be more randomly distributed, addressing the predictability problem seen in earlier schemes.

Table 1: Theoretical password space for different length of text passwords.

Number of Characters	n	Password Space (in bits)
95	6	39
95	8	53
95	10	66

PCCP is stronger against password-guessing attacks than other click-based password systems and also maintains login times and success rates comparable to text passwords [5]. However, to be seriously considered as a replacement for text passwords, PCCP needs to be at least as secure as standard text passwords. We can adjust the security of PCCP by manipulating several parameters, which in turn affect the size of the theoretical password space. However, no study of a click-based graphical password system has ever made these manipulations.

The *theoretical password space* for a password system is the total number of passwords that could be generated according to the system specifications. Ideally, a larger theoretical password space lowers the likelihood that any particular password may be guessed. For text passwords, the theoretical password space is typically reported as 95^n, where n is the length of the password, and 95 is the number of typeable characters on the US English keyboard. For PCCP, the theoretical password space is calculated as: $((w \times h)/t^2)^c$, where the size of the image in pixels ($w \times h$) is divided by the size of a tolerance square (t^2, typically 19^2), to get the total number of tolerance squares per image, then is raised to the power of the number of click-points (c). Table 1 shows the theoretical password space for several lengths of text passwords and Table 2 shows the theoretical password space for PCCP with different parameters. As shown in the tables, the theoretical password space for PCCP can be adjusted to approximate the space of text passwords of varying lengths. Password spaces grow exponentially and so are commonly represented logarithmically in terms of base-2 and reported in *bits*. For example, an 8-character text password has approximately the same password space (2^{53} or 53 bits) as a PCCP password with a small image size (451×331 pixels) and 6 click-points, or a large image size (800×600 pixels) and 5 click-points.

The *effective password space* represents the set of passwords that users are likely to create. For example, in the absence of enforced rules, users of text passwords tend to include only lowercase letters, limiting the effective password

space to 26^n. For an 8-character password, this would result in a password space of 38 bits. Only rough estimates of the effective password space are available because user choice is based on personal preference rather than mathematical principles. Commonly available text password attack tools such as *John the Ripper* [12] include dictionaries of up to 40 million entries, or 25 bits. Hotspots and patterns reduce the effective password space in click-based graphical passwords. Since PCCP significantly reduces the occurrence of hotspots and patterns, its effective password space approaches the theoretical password space. By matching the theoretical password space of PCCP to that of text passwords, the corresponding effective password space of PCCP is at least as large (and likely larger) than for text passwords.

Multiple passwords create an important issue in authentication. Users typically have many different accounts and are asked to remember many different passwords [18]. This places an increased memory burden on users, and can lead to security and usability problems such as forgetting passwords, and confusing passwords across accounts [19]. Remembering a password for one account can disrupt the memory of a password for another account. This psychological phenomenon is known as *interference* [1]. Several studies explore the impact of password interference [17, 25, 34, 7].

3. STUDY

Our study was designed to explore ways of increasing the password space of PCCP by changing the configuration of the system. With PCCP, three parameters that can be manipulated are the image size, the number of click-points per password and the size of the tolerance square. In this study, we increased the number of click-points in each password and increased the size of the images presented. Our goal was to determine which manipulation resulted in better usability and memorability for approximately equivalent password spaces (as a proxy for security). In this study, we chose to keep the size of the tolerance square constant (set to 19×19 as in previous studies) because its size is constrained by human visual acuity[16] and fine motor control.

We had two hypotheses:

Hypothesis 1: Increasing the number of click-points will decrease usability (as defined below), and increasing the size of the image will not affect usability.

Hypothesis 2: For conditions with approximately comparable theoretical password spaces, the condition with the larger image size will have better usability.

Our independent variables are the image size and the number of click-points. As shown in Table 2, there were six experimental conditions: *S5* (small image, 5 click-points); *S6* (small image, 6 click-points); *S7* (small image, 7 click-points); *L5* (large image, 5 click-points); *L6* (large image, 6 click-points); and *L7* (large image, 7 click-points). The small image size was 451×331 (the size used in the original PCCP studies [5]) and the large image size was 800×600 (standardizing to a 4:3 aspect ratio). These specific settings were chosen to approximate the theoretical password space of text passwords (compare Table 1 to Table 2). Our dependent variables concerned usability, and were success rates, duration of password entry, and number of errors. Conditions with shorter durations, fewer errors and higher success rates were judged to have better usability. The level of security was based on the theoretical password space as determined by the independent variables. We also intended to explore the effects of the different conditions on user behaviour in click-point selection, possibly resulting in clustering which reduces the effective password space.

Table 2: Configuration of the system for the different experimental conditions and distribution of participants (N).

	w	h	Click-points	Condition Name	Password Space (in bits)	N
Small	451	331	5	S5	44	14
	451	331	6	S6	53	14
	451	331	7	S7	61	14
Large	800	600	5	L5	52	14
	800	600	6	L6	63	13
	800	600	7	L7	73	14

For the first hypothesis, we compared the effects of image size and number of click-points. We hypothesized that conditions with higher numbers of click-points would have decreased usability, and conditions with larger image sizes would have similar usability to conditions with smaller image sizes. For the second hypothesis, two comparisons were made between conditions with comparable theoretical password spaces: one between S6 and L5, and one between S7 and L6. We hypothesized that S6 would have lower usability than L5, and S7 would have lower usability than L6.

We hypothesized that conditions with more click-points would have lower usability because we speculated that the cognitive load and the physical task of entering another click-point would dominate the inspection task of finding a click-point on a larger image.

A between-subjects design was used, and the 83 participants (48 females and 35 males) were randomly assigned to one of six study conditions. All participants were regular computer users accustomed to using text passwords. The majority of the participants were university undergraduates, but no participants were studying computer security.

Our study had two sessions scheduled approximately two weeks apart. Based on previous data, we anticipated that users would be very successful at remembering their passwords during the first session. We had participants wait two weeks before the second session in an effort to counteract ceiling effects and provide measurable differences.

In the first session, participants initially practiced creating and re-entering passwords for two fictitious accounts, a blog and an online gaming account. This was used to explain the experimental process and familiarize participants with the system. The practice data was discarded and participants did not need to remember these passwords later on. Next, participants created and re-entered PCCP passwords for six fictitious accounts (library, email, bank, online dating, instant messenger, and work). In the second session, participants tried to re-enter these same passwords.

There were five experiment phases over the two sessions. In the first session, participants completed the create, confirm, login and recall-1 phases. In the second session, participants completed the recall-2 phase, and were debriefed and compensated for their time. Descriptions of the experiment phases are given below. For each of the six accounts:

Create Phase: Participants select points on images to create their password.

Confirm Phase: Participants re-enter the same password to make sure they remember it. They could re-try as many times as necessary and could reset their password if it was forgotten.

Login Phase: Participants try to log in to the account using the same password. They could re-try as many times as necessary and could reset their password if it was forgotten.

Recall-1 Phase: Participants try to log in to each of their accounts in a shuffled order. Multiple attempts were allowed and participants had the option of saying they had forgotten a password to move to the next account.

Recall-2 Phase: Two weeks later, participants try to log in to their accounts in the same shuffled order. Multiple attempts were allowed and participants had the option of saying they had forgotten a password to move to the next account.

The experiment used a custom stand-alone J# application running on a Windows desktop computer. A set of 465 images was used, and no images were repeated between or within passwords for a given user. The small and large image conditions shared the same images except that they were displayed at different resolutions. Figure 1 shows the user interface for creating passwords with the two different image sizes. The size of the viewport during password creation was kept consistent at 75×75 pixels across all conditions. Similarly, the tolerance square during all password re-entry phases was 19×19 for all conditions.

Figure 1: User interface for password creation for the small and large image sizes in PCCP.

4. RESULTS

In this section, we report on the effects of the dependent variables (success rates, times and errors) of the six study conditions on success rates, errors and durations of password entry. We used statistical analysis to determine whether differences in the data were likely to reflect actual differences between conditions or whether these might reasonably have occurred by chance. Specific tests will be described throughout the section as they are reported. In all cases, we regard a value of $p < .05$ as indicating statistical significance. In such cases there is less than a 5% chance that these results occurred by chance. In the tables reporting statistics, results in bold are statistically significant.

Several figures in this section show boxplots to illustrate distributions. Boxplots show the median, the inner quartiles (as a box), the outer quartiles (as whiskers) and any outliers (as circles). The notches indicate the 95% confidence interval around the median. In some cases, the boxplot confidence intervals suggest a difference between distributions, however, the corresponding statistical test reveals no significant difference between means. In our study, we rely on the statistical tests to determine significance.

4.1 Success Rates

We report success rates at three different levels: first time success, success within three attempts, and eventual success. First time success occurs when the password is entered correctly on the first attempt, with no mistakes or restarts. Success rates within three attempts indicate that fewer than three mistakes or restarts occurred and eventual success rates indicate that the participant made multiple attempts, but was eventually successful. Success was recorded for the login, recall-1 and recall-2 phases. Chi-squared (χ^2) tests were used to examine significance since success results are non-ordered, categorical data. For clarity, successes are presented as percentages, but the statistical tests were based on contingency tables for the number of successful and failed password entries.

Hypothesis 1: We first address the login and recall-1 phases from the first session. As shown in Table 3, success rates were very high in the login and recall-1 phases. The first attempt success rates ranged between 83% and 94% for the login phase, and between 82% and 93% for the recall-1 phase, showing that participants were very successful at remembering their passwords after a short time period.

For login and recall-1, which occurred during the first session, we consider success on first attempt to be the most appropriate measure of success since users' memory of the password will still be fresh. Results of χ^2 tests for the effects of image size and number of click-points on success rates are presented in Table 4. Neither the login or recall-1 phases showed any significant effects for the number of click-points or the image size. For example, in the login phase $\chi^2(1, n = 462) = 1.335, p = 0.248$ shows no significant effect for image size on first attempt. In fact, further

Table 3: Success rates on first attempt, within 3 attempts and multiple attempts (eventual success) per phase.

	First Attempt			Within 3 Attempts			Eventual Success		
Condition	Login	Recall-1	Recall-2	Login	Recall-1	Recall-2	Login	Recall-1	Recall-2
S5	91%	87%	25%	100%	95%	37%	100%	99%	42%
S6	83%	89%	28%	99%	93%	40%	100%	93%	48%
S7	92%	85%	18%	99%	91%	32%	100%	96%	42%
L5	91%	82%	18%	100%	94%	33%	100%	94%	45%
L6	94%	93%	18%	98%	97%	27%	100%	100%	36%
L7	92%	82%	5%	100%	96%	14%	100%	100%	36%

tests found no significance for any of the three measures of success (first attempt, within 3 tries, and eventual success) in the login and recall-1 phases.

For recall-2, which occurred after two weeks, we consider success within 3 attempts as the most appropriate measure since it most closely reflects account lockout practices for real systems. In the recall-2 phase, there were significant effects of both number of click-points and image size when considering success within three attempts. Participants were more successful at logging in within three attempts when they had fewer click-points ($\chi^2(2, n = 462) = 6.663, p = 0.036$) and when the image size was smaller ($\chi^2(1, n = 462) = 6.605, p = 0.010$).

We had hypothesized that the number of click-points would have an impact and image size would not, but in fact we have evidence that both parameters affect memorability.

Hypothesis 2: We saw comparable success rates for S6 and L5 in all phases (Table 3), and statistical tests showed no significant differences. We again saw similar success rates for S7 and L6 in all phases and there were no significant differences in success rates for login, recall-1 and recall-2. This means that, for equivalent password spaces, we have no evidence that either having a larger image or more click-points affected participants' ability to more successfully remember their passwords. Therefore, the success rates offer no evidence to support hypothesis two.

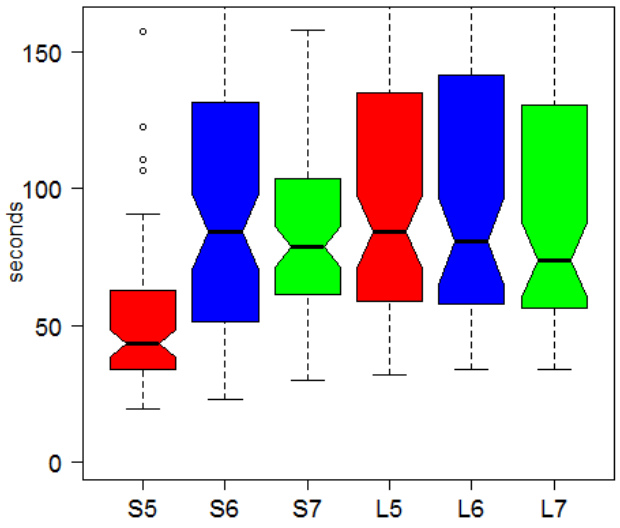

Figure 2: Create phase: boxplots showing time in seconds for each condition.

4.2 Times

Times were measured for each password entry from when the first image appeared on the screen until the participant successfully logged in to the account. This included the time to enter their username, as well as any time making errors (pressing the login button and having the system say that the password is incorrect) or resulting from restarts (analogous to pressing the backspace key when entering a text password). All password attempts that were eventually successful were included in the time calculations. We ran two-way ANOVAs to examine the main effects of number of click-points and image size. ANOVAs compare variance of the means for multiple samples and identify whether any of the samples are likely to come from different distributions.

Hypothesis 1: As shown in Table 5 and Figures 2 to 6, the number of click-points affected the time taken to create and re-enter passwords in all phases except recall-1. The image size affected the time taken to create and re-enter passwords in all phases except the login and recall-1 phases.

Although there are statistically significant differences in the times taken to create passwords, no clear pattern emerges. During the password re-entry phases (confirm, login, recall-1, recall-2), a general increase in median times can be seen in the figures as more click-points or larger images are used. These are statistically significant for the confirm phase. While visually it appears that the same is true for login and recall-1, statistical tests show no significance in these cases.

After two weeks, participants took much longer in all conditions than in session 1 to re-enter their passwords. Recall-2 times show statistically significant results, however, this trend is not as evident in Figure 6. In summary, we found evidence that both number of click-points and image size affected the time to create and re-enter passwords.

Hypothesis 2: In this section, we report on our pairs of conditions with comparable theoretical password spaces (S6 vs. L5, and S7 vs. L6). We used independent samples *t-tests* to test for significant differences in times. These tests compare variance of the means between two distributions.

S6/L5: The time distributions are available in Figures 2 to 6 and one can compare S6 with L5. As shown in Table 6, the only significant difference in durations for S6 vs. L5 was in the recall-2 phase (t(106)=-2.604, p=0.011). Participants using the system under L5 took longer than those using S6. While the differences between the distributions are statistically significant, there is only a 1.5 second difference in

Table 4: Significance tests for success rates for each phase, only the most relevant measure is reported.

	First Attempt		Within 3 Attempts
	Login	Recall-1	Recall-2
Number of Click-points	$\chi^2(2, n = 462) = 1.517$ $p = 0.468$	$\chi^2(2, n = 462) = 3.841$ $p = 0.147$	$\mathbf{\chi^2(2, n = 462) = 6.663}$ $\mathbf{p = 0.036}$
Image Size	$\chi^2(1, n = 462) = 1.335$ $p = 0.248$	$\chi^2(1, n = 462) = 0.111$ $p = 0.739$	$\mathbf{\chi^2(1, n = 462) = 6.605}$ $\mathbf{p = 0.010}$

Table 5: Mean times in seconds and two-way ANOVA results comparing all 6 conditions for each phase.

Condition	Create	Confirm	Login	Recall-1	Recall-2
S5	66.8	20.8	15.9	20.7	43.5
S6	108.7	23.2	19.5	21.0	60.6
S7	82.4	28.1	20.7	24.2	65.5
L5	106.9	24.1	18.0	18.7	62.1
L6	103.7	30.3	20.9	23.7	76.4
L7	93.4	31.6	22.0	27.8	89.5
Number of Click-points	$\mathbf{F(2, 456) = 4.55}$, $\mathbf{p = 0.011}$	$\mathbf{F(2, 456) = 8.48}$ $\mathbf{p = 0.000}$	$\mathbf{F(2, 456) = 10.87}$ $\mathbf{p = 0.000}$	$F(2, 456) = 2.92$ $p = 0.055$	$\mathbf{F(2, 456) = 4.12}$ $\mathbf{p = 0.018}$
Image Size	$\mathbf{F(1, 456) = 6.87}$ $\mathbf{p = 0.009}$	$\mathbf{F(1, 456) = 9.83}$ $\mathbf{p = 0.002}$	$F(1, 456) = 3.81$ $p = 0.052$	$F(1, 456) = 0.43$ $p = 0.513$	$\mathbf{F(1, 456) = 8.33}$ $\mathbf{p = 0.004}$

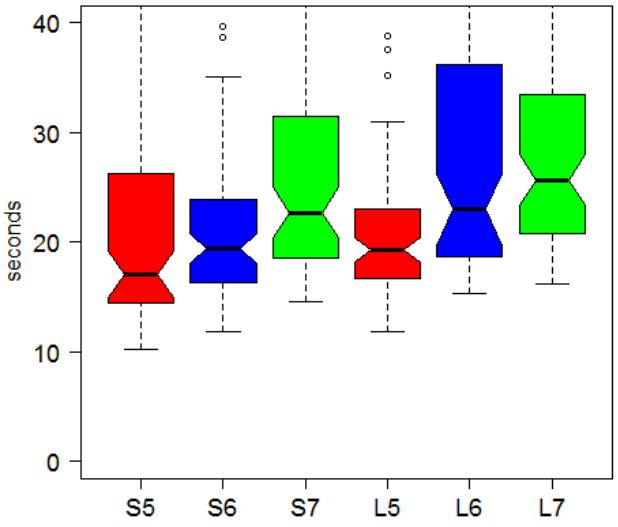

Figure 3: Confirm phase: boxplots showing time in seconds for each condition.

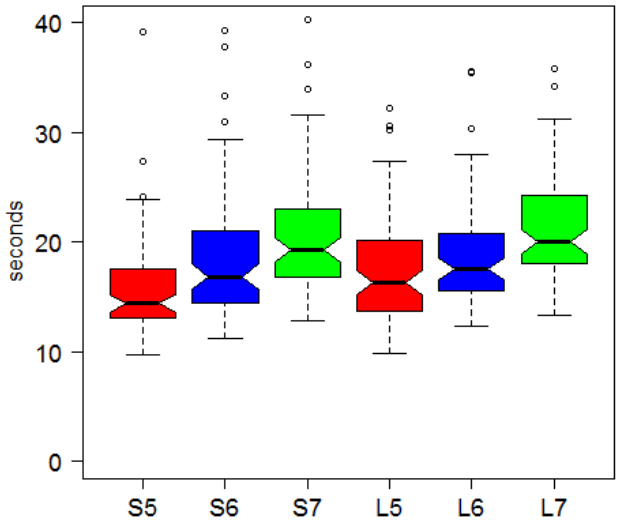

Figure 4: Login phase: boxplots showing time in seconds for each condition.

the mean times, indicating only a small difference in practice. Therefore, this provides little evidence to support our hypothesis 2 with respect to times.

S7/L6: For S7 and L6, a significant difference was seen in create times ($t(98) = -2.654$, $p = 0.009$). Password creation in L6 took longer on average than S7, indicating that participants took longer selecting click-points on the larger image as opposed to choosing an additional click-point on smaller images. As shown in Table 6, no significant differences were seen in any of the password re-entry phases. For example, in the recall-2 phase, a t-test found $t(146) = 0.114, p = 0.910$. Once passwords were created, participants could recall them equally quickly in both the S7 and L6 conditions. Figures 2 to 6 show the time distributions.

We found little evidence to support hypothesis 2. In fact, the only significant results (S6 vs. L5 recall-2 password entry and L6 vs. S7 password creation time) showed evidence contrary to this hypothesis since L6 took longer than S7.

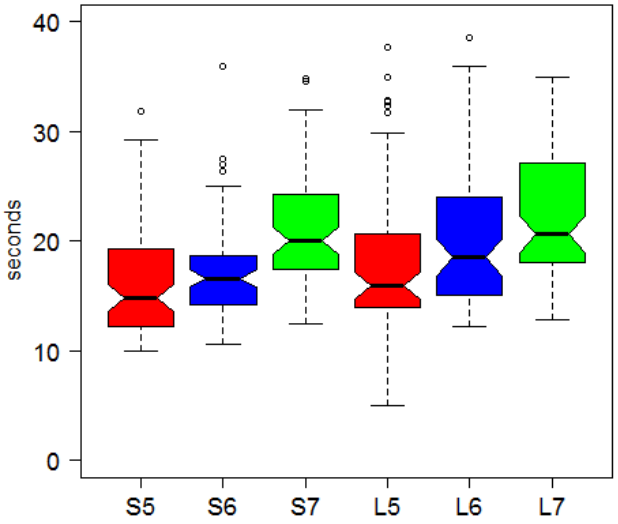

Figure 5: Recall-1 phase: boxplots showing time in seconds for each condition.

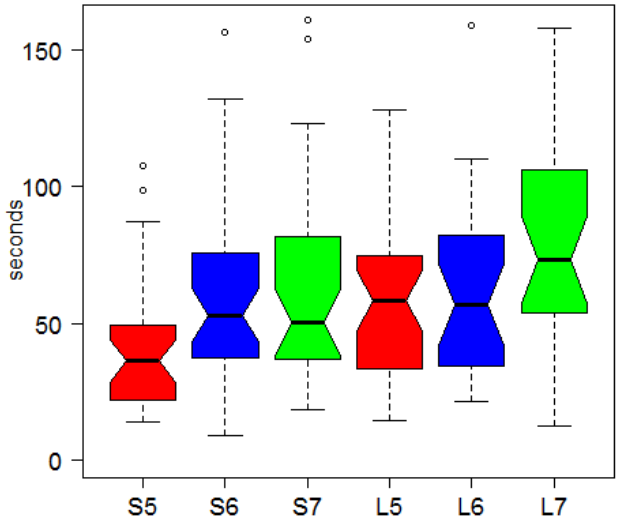

Figure 6: Recall-2 phase: boxplots showing time in seconds for each condition.

Table 6: t-tests for durations: Hypothesis 2

Phase	S6 vs. L5	S7 vs. L6
Create	$t(156) = 0.145$ $p = 0.885$	$\mathbf{t(98) = -2.654}$ $\mathbf{p = 0.009}$
Confirm	$t(146) = -0.412$ $p = 0.681$	$t(136) = -0.780$ $p = 0.437$
Login	$t(148) = 1.157$ $p = 0.249$	$t(89) = -0.078$ $p = 0.938$
Recall-1	$t(127) = 0.165$ $p = 0.869$	$t(120) = 0.327$ $p = 0.744$
Recall-2	$\mathbf{t(106) = -2.604}$ $\mathbf{p = 0.011}$	$t(146) = 0.114$ $p = 0.910$

4.3 Errors

An error was recorded every time a participant chose the wrong click-point or restarted their password attempt. Since error distributions were non-normal, we used several non-parametric tests for analysis. When comparing across all conditions, we ran Kruskal-Wallis tests, which are similar to ANOVAs, but used when the distribution of the samples is skewed, as is common with error counts. When comparing two specific conditions, we conducted Wilcoxon tests (also known as Mann-Whitney tests) to check for significant differences. Wilcoxon tests are similar to independent sample t-tests, but make no assumptions about the distributions of the compared samples.

Hypothesis 1: Participants in all conditions made very few errors when entering their passwords after a short time interval. For the confirm, login and recall-1 phases, the mean number of errors per account for each phase was less than 1 (Table 7). Participants made many more errors in the recall-2 phase, with means ranging between 0.97 and 4.70 across the 6 conditions.

Table 7: Mean number of errors per phase.

Condition	Confirm	Login	Recall-1	Recall-2
S5	0.39	0.14	0.42	0.97
S6	0.28	0.28	0.05	1.13
S7	0.33	0.11	0.29	1.82
L5	0.45	0.10	0.12	1.77
L6	0.34	0.10	0.16	3.44
L7	0.64	0.09	0.43	4.70

As in the duration analysis, we created three distributions to explore the effects of number of click-points, aggregating the image sizes in each. To explore the effect of image size, we created two distributions, aggregating the number of click-points in each. In both cases we used the Kruskal-Wallis statistic (conventionally reported as χ^2).

For confirm and login, neither increasing the number of click-points nor image size had any significant effect on errors. In the recall-1 phase we saw significant differences between the three distributions based on the number of click-points, but no clear pattern emerged. We did not see any significant differences between image sizes. In the recall-2 phase, there was a significant effect of both click-points ($\chi^2(2, n = 192) = 11.765, p = 0.003$) and image size ($\chi^2(1, n = 192) = 10.260, p = 0.001$). This indicated that having more click-points or larger images caused participants to make more errors after two weeks.

We found no evidence to support hypothesis 1 with respect to number of errors because significant effects were found for both number of click-points and image size. In fact, the significant result for number of click-points in the recall-1 phase contradicts our hypothesis.

Hypothesis 2: For S6 and L5, the mean number of errors was less than 1 for the confirm, login and recall-1 phases

(Table 7). In the recall-2 phase, the mean number of errors for S6 was 1.13 and for L5 was 1.77. Wilcoxon tests found no significant differences in any phases.

There were few differences between S7 and L6. In the confirm, login and recall-1 phases, the mean number of errors for both conditions was also less than 1. As expected, participants made more errors in the recall-2 phase than in other phases (on average, 1.82 and 3.44 errors per account for S7 and L6 respectively). However, none of these differences were statistically significant, providing no evidence that having a larger image or more click-points make different demands on usability.

Analysis of the number of errors does not provide any evidence to support hypothesis 2. We found no significant differences in errors between any of the conditions with comparable theoretical password spaces.

4.4 Summary of Results

We chose three measures of usability: success rates, durations and number of errors. Summaries of the results for hypothesis 1 and hypothesis 2 are available in Tables 8 and 9. As we describe above, phases from the first session (create, confirm, login, and recall-1) use success on first attempt as the measure of success. Recall-2 uses success within 3-attempts instead. Times and errors include all activity until successful login.

Hypothesis 1: Hypothesis 1 suggests that increasing the number of click-points will decrease usability, but increasing the size of the image will not affect usability.

We found little support for our hypothesis because any significant differences were either contradictory or apparent in both number of click-points and image size. More errors were made when participants had an increased number of click-points, but not when they had larger images. Over the short-term, statistically significant differences are found in the times taken to enter passwords. However, these differences affect both the number of click-points and image sizes, therefore do not support hypothesis 1.

After 2 weeks, statistically significant differences were found in each of the measures reported. There is no clear pattern for success rates, but durations and number of errors increase as the number of click-points or image size increases. Again, this contradicts hypothesis 1.

Hypothesis 2: Hypothesis 2 suggests that for conditions with approximately comparable theoretical password spaces, the condition with the larger image size will have better usability.

Our results provide little support for this hypothesis. For success rates, we saw no differences in any of the phases between conditions with similar password spaces. For times, we found a few differences between conditions but these were contradictory to our hypothesis. Finally, we found no significant differences between conditions with respect to errors. Overall, our experiment results do not support our hypothesis that conditions with larger image sizes would have better usability.

5. CLICK-POINT CLUSTERING

During PCCP password creation, users pressed the shuffle button when they were unable or unwilling to select a click-point within the currently highlighted viewport. In terms of security, we expect fewer shuffles to lead to more randomly distributed passwords. Figure 7 shows the distribution of

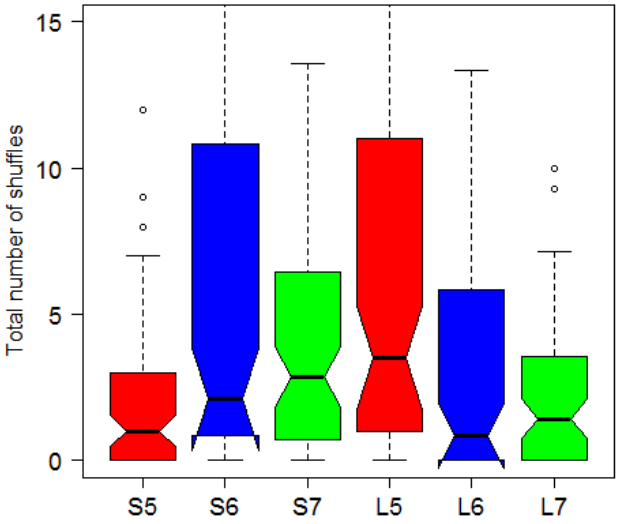

Figure 7: Boxplot of total shuffles per account by condition

shuffle counts for each condition. Large variability is seen, but no clear pattern emerges. The median number of shuffles per password for all conditions is less than five, indicating that most participants pressed the shuffle button less than once per image (passwords consisted of between 5 and 7 images).

Passwords should be as random as possible while still maintaining memorability. Clustering of click-points on an image across users creates what are known as *hotspots*. Attackers who can determine likely hotspots (through image analysis or by gathering a sample of passwords from even a small number of people [32]) would be better positioned to launch an effective dictionary guessing attack. Ideally, a system would minimize the occurrence of hotspots. PCCP attempts to accomplish this through the randomly-positioned viewport, however, users may shuffle the viewport to find a memorable location. We explored whether either image size or number of click-points had an effect on user choice.

To analyze the randomness and clustering of our two-dimensional spatial data, we turned to point pattern analysis [14] commonly used in biology and earth sciences. Our analysis used *spatstat* [2], a spatial statistics package for the *R* programming language.

We used the *J-statistic* [33] as a measure of click-point clustering on a subset of images for which we had sufficient data. Our system ensured that every participant saw 30 of the images, giving enough data points for analysis. To measure the clustering of points in a dataset, the J-statistic combines nearest-neighbour calculations and empty-space measures for a given radius r. When $J(r) = 0$, it indicates that all points cluster at the same location. When $J(r) = 1$, the points are randomly dispersed across the space. Finally, when $J(r) > 1$, the points are uniformly distributed. For passwords, we want results closer to $J(r) = 1$ since this would be least predictable by attackers. We examined clus-

Table 8: Results summary for hypothesis 1: checkmarks represent statistically significant results.

	Success Rates		Times		Errors	
	Image Size	Click-Points	Image Size	Click-points	Image Size	Click-points
Create			✓	✓		
Confirm			✓	✓		
Login				✓		
Recall-1						✓
Recall-2	✓	✓	✓	✓	✓	✓

Table 9: Results summary for hypothesis 2: checkmarks represent statistically significant results.

	Success Rates		Times		Errors	
	S6 vs. L5	S7 vs. L6	S6 vs. L5	S7 vs. L6	S6 vs. L5	S7 vs. L6
Create				✓		
Confirm						
Login						
Recall-1						
Recall-2			✓			

tering at $J(9)$. A radius of 9 approximates the size of the 19×19 tolerance squares used by our system during password re-entry.

Figure 8 shows the level of clustering for the 30 images, with the image names on the x-axis. This figure illustrates the effects of the number of click-points on clustering. Points on each line contain statistics for passwords created using either 5, 6, or 7 click-points. The J-statistic for each image is distinct; the connecting lines are only included for readability. For each of the three cases, data from the small (451×331) and large (800×600) images are grouped together based on the number of click-points per password. For example, the 5 click-point line represents all passwords containing 5 click-points regardless of whether they were created on small or large images. The point coordinates on the large images are re-scaled to the coordinate system of the small image so that all data is presented at 451×331 dimensions. This aligns features on the small and large versions of the same images.

The boxplot presented on the right of Figure 8 summarizes the distribution of the statistics for the three cases. This aggregation is included to better illustrate the range of values represented by each set of J-statistics. The lines on the main graph do not show any consistent relationship between each other. The boxplot shows overlapping ranges and no consistent relationship as the number of click-points increases.

To our knowledge, there is no statistical test to compare sets of J-statistics to each other. If we regard the data as categorical, we can identify six categories stemming from the possible orderings: 567, 576, 657, 675, 756, 765. For example, in Figure 8 the *alphamat* image falls in the 576 category because J(9) for 5 click-points is larger than J(9) for 7 click-points, which is larger than J(9) for 6 click-points. We can then apply a chi-squared test between the observed results and the expected results (equal probability for each category). This test shows no significant differences ($\chi^2(5, n = 60) = 5.675, p = 0.339$). We therefore find no evidence for a difference in clustering between the different numbers of click-points.

Figure 9 shows the level of clustering for the 30 images, distinguishing the effects of image size. Each line contains the statistics for passwords created on either the small or large images. For each of the two cases, data from 5, 6, and 7 click-points are combined. In other words, all passwords created on large images (regardless of how many click-points) are grouped together, and all passwords created on small images (regardless of how many click-points) form a second group. The data from the large images are again scaled to ensure comparability of the J-statistic.

To the right of Figure 9, a boxplot summarizes the distribution of the statistics for both cases. For most images, the main graph indicates that the larger images have less clustering (J(9) closer to 1) than the smaller images. The boxplot also suggests that the distributions are distinct. If we regard the data as categorical, we could distinguish two representing whether the small or large image size has stronger clustering. We applied a chi-squared test between the observed results and the expected results (equal probability for each category). This test shows a significant difference in clustering for the small and large images ($\chi^2(1, n = 60) = 9.603, p = 0.002$), indicating that larger images have significantly less clustering.

In summary, from Figure 8 it appears that additional click-points do not lead to user behaviour resulting in more clustering. However, larger images appear to influence user choice towards less clustering. This result suggests that PCCP's viewport and shuffle mechanism is more effective in reducing clustering, and therefore promoting security, when used with larger images.

6. DISCUSSION

6.1 Interpretation of Results

Contrary to our expectations, we did not see any large differences in how the number of click-points and image size affected usability. We expected that increasing the image size would have little or no effect on usability and memorability but we found that it had a similar effect to increasing the number of click-points.

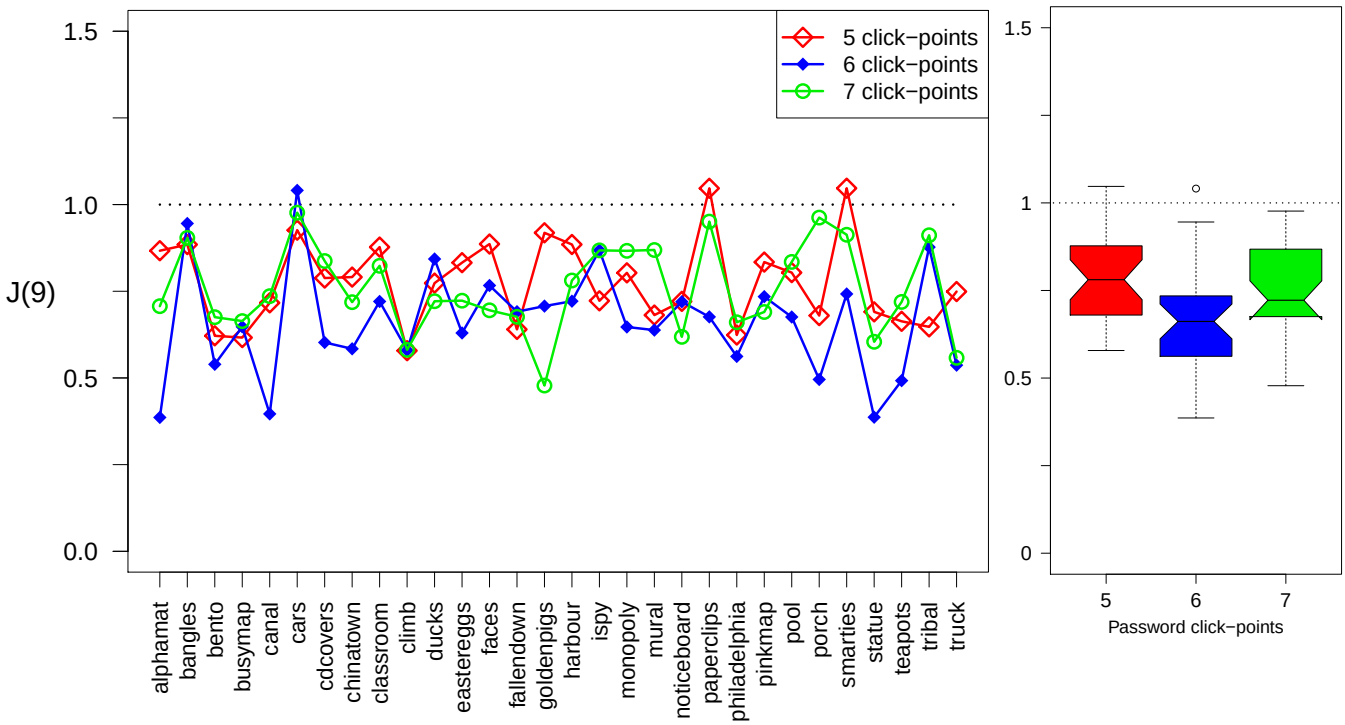

Figure 8: J-statistics for distributions of 5, 6, or 7 click-points. Data from the larger image is scaled to allow for aggregation.

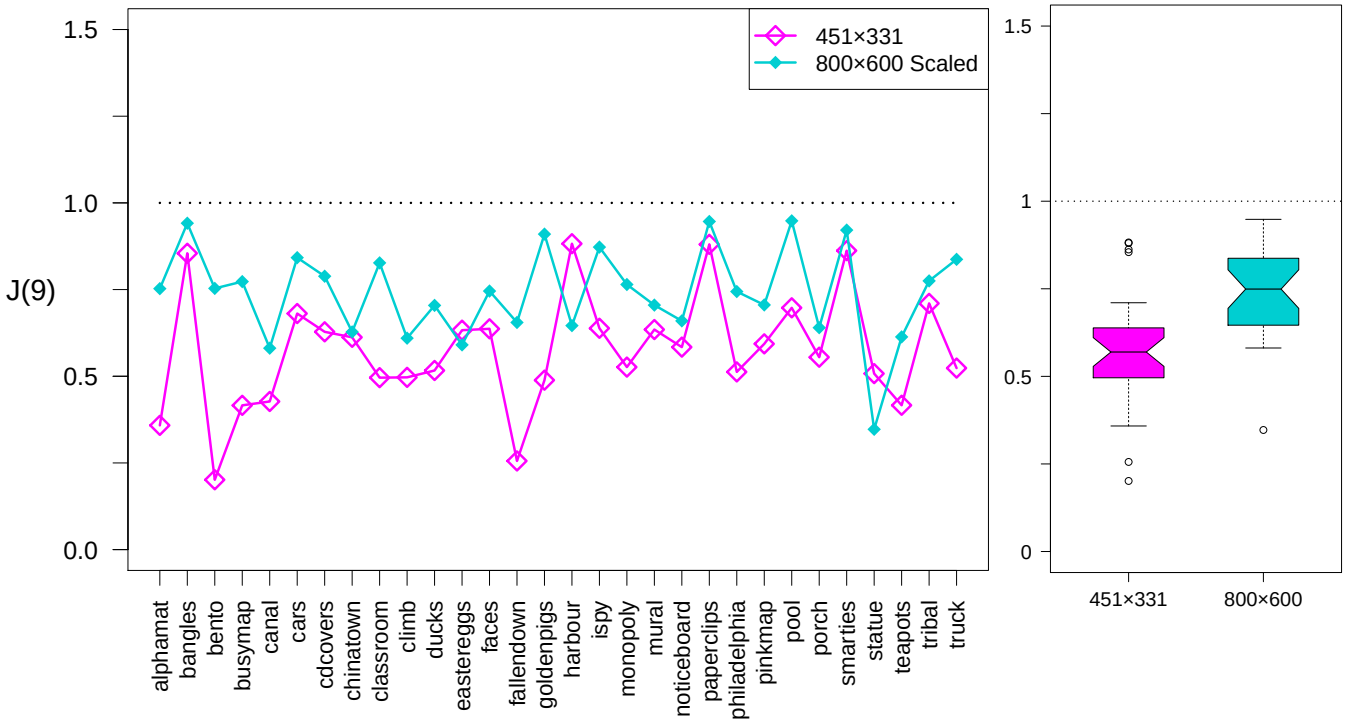

Figure 9: J-statistic for distributions of small and large images. Data from the larger images is scaled to allow generation of comparable J-statistics.

The lack of difference in usability between increasing the number of click-points and image size presents an opportunity, suggesting that other considerations can be taken into account when increasing security. In a situation where choosing a click-point was comparatively difficult (as for a person with a poor fine motor control), this might be accommodated by having fewer click-points, but larger images. More click-points might be appropriate in a situation where screen size was limited, such as on a mobile device. The equivalent demands on usability when increasing security thus give increased flexibility in designing.

The conditions under which participants created and used their passwords are clearly artificial. In real life, it is extremely unlikely that a user would create six passwords in a row, then not see them again for two weeks, until they tried to log into all six accounts. The design of our study was meant to emphasize differences between the six conditions by making the task harder. The results of the study for the create, confirm, and login phases are similar to results seen in an earlier study of PCCP [5] and are consistently good, with only small differences between conditions. Further work is needed to confirm real-life usability. We have developed a web-based infrastructure that will allow us to conduct such tests in the near future.

Another issue to consider in our lab study is password interference. Participants each created six passwords, each of which was only tenuously linked to a user account. These accounts (library, email, bank, blog, online dating, instant messenger, and work) were denoted only by coloured banners on the login screen (see Figure 1). Although we attempted to emphasize to the user that each account was distinct, there was no practical difference between them. In real life, accounts would be separated from each other by appearance of the website, or created at different times. Participants likely had a hard time distinguishing their passwords from each other, and this might have led to more difficulty in remembering them after two weeks.

6.2 Principles and Implications

Although our study focused on several specific configurations of PCCP, it is important to consider the general underlying principles involved. In PCCP there are three obvious parameters that can be changed to increase security without changing the nature of the scheme: image size, number of clicks, and the tolerance region. Decreasing the tolerance region is constrained by the visual acuity and motor precision of the user, and we chose not to investigate those issues.

Image Size: The size of the images shown in each password seems to relate to several human factors. The user likely responds to the appearance of the image with a quick visual survey of the image. While principles of visual attention apply to this survey, the nature of the survey may change with familiarity, or even with exposure to other images or events that relate to the image. The human visual system involves several approaches, including taking in the overall impression, and responding to various attractors. Our initial speculation was that these might be the dominant factors, and we did not expect them to vary much with image size.

For closer inspection of an image, however, the eye will be directed to specific parts of the image. Such close visual inspection requires high acuity vision using the fovea, the area of the retina with a high density of photoreceptor cells[16]. The size of the fovea limits foveal vision to an angle of approximately 1° within the direct line to the target of interest. At a normal viewing distance for a computer screen, say $60cm$, this results in sharp vision over an area of approximately $4cm^2$. The size of the image, and the number of attractors, will then determine the number of foveal areas the user will inspect, and the distance of the saccades as they move from one target to another will also be a factor.

Several factors will affect how a PCCP user surveys an image. PCCP is a cued-recall scheme, so the user will be looking for cues to remind them where to click. PCCP also gives implicit feedback with each image about the previous click, by displaying the correct image if the user chooses the correct click-point. This means that the user will be assessing whether or not the current image is familiar to them. Then, once the user has recognized the image and found their click-point, they must position the cursor correctly using a mouse, touchpad or other pointing device. The time taken to position the cursor may be predicted by Fitts' Law, which determines targeting time from the distance and target size [24]. However, we typically observe users moving the cursor to follow their gaze as they examine the image, so the final movement to a click-point is typically very short.

Click-points: The number of click-points in a PCCP password involves a repetition of all the elements involved in finding and clicking on a single point. We initially assumed this repetition would make the number of click-points a more important factor than the size of the image in determining the usability, but the study results did not support this. In a pure-recall system, we would expect to see serial memory effects, which cause people to better remember the items at the beginning and end of an ordered list. With PCCP's cued-recall, however, we expect milder serial memory effects, because participants respond to each picture as an individual cue. However, it is certainly possible that users begin to learn the pattern of click-points and anticipate where to focus their gaze, and move their cursor. This anticipation may reduce the work needed per image in ways that have not yet been fully explored.

Alternative Configurations: It appears that factors such as increasing the number of click-points or image size balance each other out, at least for the settings in our study. To consider the general underlying principles, we might speculate about more extreme possibilities. In our study, the two image sizes used were 451×331 pixels and 800×600 pixels. The tolerance region of the scheme was 19×19 pixels, which meant that the images had approximately 414 and 1330 click areas distinguishable to the system, respectively. The LCD display we used measured $43cm$ ($17in$) diagonally with a resolution of 1280×1024 pixels. The small image measured about $12cm \times 9cm$, or $84cm^2$, and the large image about $21cm \times 16cm$ or $336cm^2$. Our study showed that users can cope with inspecting and selecting click-points on images of both sizes within a reasonable amount of time: mean login time of approximately 20 seconds, including entry of username and all click-points.

In our S6 and L5 conditions, the theoretical password space is about 52 bits. In S7 and L6, it is about 62 bits. Knowing that the image sizes in these conditions were usable, we explore larger sizes in order to decrease the number of click-points while keeping the password space the same. Table 10 shows some possibilities. For example, even requiring only 3 clicks and keeping the aspect ratio the same would

require an image size of 8916 × 6687 pixels for 52 bits, and 28305 × 21229 pixels for 62 bits. These would seem to be unreasonable sizes for graphical password images, and would involve a very large number of areas to be inspected. As the number of click-point required decreases, the size of the images implied must grow exponentially, and quickly reaches the bounds of usability. We do navigate on very large *virtual* displays when using cartographic browsers such as Google Earth. This is only manageable, however, through the use of the zoom and pan capabilities, and so the interaction in fact involves a number of clicks.

Table 10: Image sizes required, based on password space and number of clicks.

Bits	Clicks	Xpixels	Ypixels	Xcm	Ycm
52	6	442	332	11	9
52	5	806	605	21	16
52	4	1986	1489	51	38
52	3	8916	6687	229	171
52	2	179727	134795	4608	3456
62	6	788	591	20	15
62	5	1613	1210	41	31
62	4	4723	3542	121	91
62	3	28305	21229	726	544
62	2	1016688	762516	26069	19552

Implications for Mobile Devices: Our participants managed well with passwords of 5, 6, and 7 click-points in length, so an alternative exploration might be to consider more click points, and allow the image size to be reduced while still maintaining a large password space. Table 11 shows possibilities, using typical small sizes on mobile devices. For example, a small mobile phone might have 120×80 pixels, whereas a Blackberry Curve 8300 has 320 × 240 pixels, while the Blackberry Bold and the Apple iPhone have 480 × 320 pixels. Mobile devices sometimes involve a touchscreen instead of a stylus, and often use a dense pixel pitch so images appear physically smaller than the equivalent dimensions on a computer screen. In the table, we accommodate this by using a tolerance region for the mobile devices of 38×38: the size of square onscreen keyboard elements on an iPhone. For the iPhone screen, this would require 8 clicks for a 52 bit password space. These numbers seem potentially acceptable, especially as we frequently type words of that many characters. This suggests that a graphical password scheme such as PCCP might be usable on mobile devices. The small screens will not be compatible with the current viewport because its current size highlights too much of the image to effectively reduce clustering. We are currently exploring a redesigned viewport mechanism. The increasing use of mobile devices for secure online transactions indicates a need for more secure passwords than simple screen unlock mechanisms, and we believe a system such as PCCP has potential for both usability and security.

7. CONCLUSION

In this paper, we explored the issue of how increasing the security of a click-based graphical password scheme would affect usability and memorability. We tested PCCP with different parameters in order to evaluate its usability when the theoretical password space is increased. We found that increasing the number of click-points or increasing the image size both have usability and memorability impacts. While varying parameters so as to hold constant the size of the theoretical password space, we found no evidence of differences between configurations varying the number of click-points and image size. Additionally, we explored the effects of number of click-points and image size on user behaviour resulting in clustering of click-points. We found no evidence that the number of click-points had an effect, but it appeared that larger images led to less clustering.

Table 11: Click numbers required, based on password space and image size.

Bits	Xpixels	Ypixels	Tolerance	Clicks
52	800	600	19	5
52	451	331	19	6
52	320	480	38	8
52	240	320	38	9
52	80	120	38	19
62	800	600	19	6
62	451	331	19	7
62	320	480	38	9
62	240	320	38	11
62	80	120	38	23

These results have important implications for practical configuration of graphical password schemes in various contexts. For example, the results suggest that for mobile devices with small screens, it should be possible to increase security by using smaller images and more click-points while retaining usability and memorability. Conversely, larger images appear to lead to less clustering, suggesting an issue that should be considered in future research.

8. ACKNOWLEDGMENTS

The second author acknowledges NSERC Postgraduate Scholarship funding for this work. The fourth author is Canada Research Chair in Internet Authentication and Computer Security, and acknowledges NSERC funding of this chair, a Discovery Grant, and a Discovery Accelerator Supplement. The fifth author acknowledges funding of an NSERC Discovery Grant. Partial funding from the NSERC Internetworked Systems Security Network (ISSNet) is also acknowledged.

9. REFERENCES

[1] M. Anderson and J. Neely. 8: Interference and inhibition in memory retrieval. In E. Bjork and R. Bjork, editors, *Memory. Handbook of Perception and Cognition*, pages 237–313. Academic Press, 1996.

[2] A. Baddeley and R. Turner. R. Spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42, 2005.

[3] K. Bicakci, M. Yuceel, B. Erdeniz, H. Gurbaslar, and N. Atalay. Graphical Passwords as Browser Extension: Implementation and Usability Study. In *Third IFIP WG 11.11 International Conference on Trust Management*, Purdue University, USA, June 2009.

[4] R. Biddle, S. Chiasson, and P. C. van Oorschot. Graphical passwords: Learning from the first generation. Technical Report TR-09-09, Computer Science, Carleton University, `http://www.scs.carleton.ca/research/tech_reports`, 2009.

[5] S. Chiasson, A. Forget, R. Biddle, and P. C. van Oorschot. Influencing users towards better passwords: Persuasive Cued Click-Points. In *Human Computer Interaction (HCI), The British Computer Society*, 2008.

[6] S. Chiasson, A. Forget, R. Biddle, and P. C. van Oorschot. User interface design affects security: Patterns in click-based graphical passwords. *International Journal of Information Security*, 8(5), 2009.

[7] S. Chiasson, A. Forget, E. Stobert, P. C. van Oorschot, and R. Biddle. Multiple password interference in text and click-based graphical passwords. In *ACM Conf. on Computer and Communications Security (CCS)*, 2009.

[8] S. Chiasson, P. C. van Oorschot, and R. Biddle. A usability study and critique of two password managers. In *15th USENIX Security Symposium*. Usenix, August 2006.

[9] R. G. Crowder and R. L. Greene. Serial Learning: Cognition and Behaviour. In E. Tulving and F. I. Craik, editors, *The Oxford Handbook of Memory*, chapter Chapter 8, pages 125–135. Oxford University Press, 2000.

[10] D. Davis, F. Monrose, and M. Reiter. On user choice in graphical password schemes. In *13th USENIX Security Symposium*, August 2004.

[11] A. De Angeli, L. Coventry, G. Johnson, and K. Renaud. Is a picture really worth a thousand words? Exploring the feasibility of graphical authentication systems. *International Journal of Human-Computer Studies*, 63(1-2):128–152, 2005.

[12] S. Designer. John the Ripper password cracker. `http://www.openwall.com/john/`.

[13] R. Dhamija and A. Perrig. Déjà Vu: A user study using images for authentication. In *9th USENIX Security Symposium*, 2000.

[14] P. Diggle. *Statistical Analysis of Spatial Point Patterns*. Academic Press: New York, NY, 1983.

[15] A. Dirik, N. Menon, and J. Birget. Modeling user choice in the Passpoints graphical password scheme. In *3rd ACM Conference on Symposium on Usable Privacy and Security (SOUPS)*, July 2007.

[16] A. Duchowski. *Eye Tracking Methodology: Theory and Practice*. Springer, 2nd edition, 2007.

[17] K. Everitt, T. Bragin, J. Fogarty, and T. Kohno. A comprehensive study of frequency, interference, and training of multiple graphical passwords. In *ACM Conference on Human Factors in Computing Systems (CHI)*, April 2009.

[18] D. Florencio and C. Herley. A large-scale study of WWW password habits. In *16th ACM International World Wide Web Conference (WWW)*, May 2007.

[19] S. Gaw and E. Felten. Password management strategies for online accounts. In *2nd Symposium On Usable Privacy and Security (SOUPS)*, July 2006.

[20] K. Golofit. Click passwords under investigation. In *12th European Symposium On Research In Computer Security (ESORICS), LNCS 4734*, September 2007.

[21] W. Jansen, S. Gavrila, V.Korolev, R. Ayers, and R. Swanstrom. Picture password: A visual login technique for mobile devices. Technical Report NISTIR 7030, National Institute of Standards and Technology, 2003.

[22] I. Jermyn, A. Mayer, F. Monrose, M. Reiter, and A. Rubin. The design and analysis of graphical passwords. In *8th USENIX Security Symposium*, August 1999.

[23] L. Jones, A. Anton, and J. Earp. Towards understanding user perceptions of authentication technologies. In *ACM Workshop on Privacy in Electric Society*, 2007.

[24] I. S. MacKenzie. Fitts' law as a research and design tool in human-computer interaction. *Hum.-Comput. Interact.*, 7(1):91–139, 1992.

[25] W. Moncur and G. Leplatre. Pictures at the ATM: Exploring the usability of multiple graphical passwords. In *ACM Conference on Human Factors in Computing Systems (CHI)*, April 2007.

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

[27] K. Renaud. Guidelines for designing graphical authentication mechanism interfaces. *International Journal of Information and Computer Security*, 3(1):60 – 85, June 2009.

[28] A. Salehi-Abari, J. Thorpe, and P. C. van Oorschot. On purely automated attacks and click-based graphical passwords. In *24th Annual Computer Security Applications Conference (ACSAC)*, 2008.

[29] M. A. Sasse, S. Brostoff, and D. Weirich. Transforming the 'weakest link' – a human/computer interaction approach to usable and effective security. *BT Technology Journal*, 19(3):122–131, July 2001.

[30] X. Suo, Y. Zhu, and G. Owen. Graphical passwords: A survey. In *Annual Computer Security Applications Conference (ACSAC)*, December 2005.

[31] H. Tao and C. Adams. Pass-Go: A proposal to improve the usability of graphical passwords. *International Journal of Network Security*, 7(2):273–292, 2008.

[32] J. Thorpe and P. C. van Oorschot. Human-seeded attacks and exploiting hot-spots in graphical passwords. In *16th USENIX Security Symposium*, August 2007.

[33] M. van Lieshout and A. Baddeley. A nonparametric measure of spatial interaction in point patterns. *Statistica Neerlandica*, 50(3):344–361, 1996.

[34] K.-P. L. Vu, R. Proctor, A. Bhargav-Spantzel, B.-L. Tai, J. Cook, and E. Schultz. Improving password security and memorability to protect personal and organizational information. *International Journal of Human-Computer Studies*, 65:744–757, 2007.

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