Order and Entropy in Picture Passwords

Saranga Komanduri*
Bowling Green State University

Dugald R. Hutchings†
Bowling Green State University

Abstract

Previous efforts involving picture-based passwords have not focused on maintaining a measurably high level of entropy. Since password systems usually allow user selection of passwords, their true entropy remains unknown. A 23-participant study was performed in which picture and character-based passwords of equal strength were randomly assigned. Memorability was tested with up to one week between sessions. The study found that both character and picture passwords of very high entropy were easily forgotten. However, when password inputs were analyzed to determine the source of input errors, serial ordering was found to be the main cause of failure. This supports a hypothesis stating that picture-password systems which do not require ordered input may produce memorable, high-entropy passwords. Input analysis produced another interesting result, that incorrect inputs by users are often duplicated. This reduces the number of distinct guesses users can make when authentication systems lock out users after a number of failed logins. A protocol for ignoring duplicate inputs is presented here. A shoulder-surfing resistant input method was also evaluated, with six out of 15 users performing an insecure behavior.

Keywords: Graphical authentication; picture superiority; shoulder surfing

Index Terms: K.6.5 [Management of Computing and Information Systems]: Security and Protection—Authentication; H.1.2 [Information Systems]: Models and Principles—User/Machine Systems; H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces;

1 Introduction

In password analysis, *entropy* is a commonly used measure of password security, where higher entropy is more desirable. Entropy can be viewed as a measure of randomness, and is typically enforced using password policies that prevent the use of dictionary words or common passwords. These policies may also require varying letter case and the insertion of numbers, since this increases entropy.[1]

However, a true measure of entropy, as defined by Shannon [21], cannot be computed without absolute knowledge of the frequency distribution of passwords. Common passwords often come from sources in popular culture [20], which suggests that the frequency distribution of passwords is continually changing. This makes both password analysis and policy enforcement difficult. For example, if an organization implements a password policy that restricts common passwords, they should regularly update their password system's dictionary to effectively screen for common passwords.

The desire for higher entropy in passwords can also produce a reduction in password usability as users struggle to remember more random passwords. This conflict between password policies and memorability has been labeled "the password problem" [31] and is central to studies that seek to improve the usability of password systems. Picture-based passwords[2] have been proposed as a solution to the password problem because pictures are more easily remembered than text. This is a phenomenon known as the *picture superiority effect*. However, the use of pictures does not, in itself, increase the entropy of user-selected passwords. Davis et al. [5] found, in the PassFaces^{TM}system, users selected attractive faces or faces of the user's own race more than others. This makes the passwords easier to guess and suggests that user-selected passwords in the PassFaces^{TM}system have undesirably low entropy. Similar results have been found for the PassPoints system [29, 9], another picture-password system [31]. These are not failures of the respective password systems but instead illustrate a general problem with user-selected passwords.

Such uncertainties in password composition can be avoided if passwords are randomly composed and assigned to users rather than user-selected. User selection of passwords is thought to be necessary for memorability. However, when pictures are employed, random assignment becomes a viable option for password systems. In studies first performed by Shepard in 1967 [22], and many others since [7, 25], participants were able to learn arbitrary images with a great deal of accuracy. These studies support the use of pictures to produce memorable, high-entropy password systems.

2 Related Work

The implementations of picture-based password systems are surprisingly varied. The earliest example appears to be that of Blonder [2] who described a system in which users "tap" points on a particular picture in a chosen sequence. This type of system has been more recently studied by Wiedenbeck et al. in their PassPoints system [32, 31].

The PassPoints system employs a static challenge/response scheme. The challenge consists of a large, photographic image, and the user responds by choosing points on the image in a particular order. If the response points are within a tolerance amount of the user's pre-selected PassPoints, the authentication is successful. This system has the potential for extremely high entropy because there are hundreds of possible memorable points in the challenge image [31].[3]

Other picture-password systems are designed for relatively low-entropy applications, such as the PIN-input systems used in ATMs. The traditional password for these systems consists of a four-digit number. In the PassFaces^{TM}system, the user must select the correct face from a grid of nine faces, and must do so four times to authenticate [3]. In the VIP system proposed by De Angeli et al. [6], the user must chose four pictures from a grid of 10 or 16. These systems are intended for comparison with standard PIN systems which have an entropy of, at most, only about 13 bits. Though it may be

*e-mail: sarangakomanduri@hotmail.com

†e-mail: drhutch@cs.bgsu.edu

[1] Entropy is a tricky term when used in password analysis, since it is not always predictably related to the expected number of guesses needed to find a password [13]. However, this topic is beyond the scope of this paper, which will follow the standard convention of using entropy as a measure of the strength of passwords.

[2] The term *password* in this paper will be used to describe both character-based and picture-based authentication systems.

[3] However, as explained earlier, studies have found common "hot spots" in user-selected PassPoints passwords [29, 9].

possible to scale such systems to handle high-entropy passwords, studies of this type have not been reported.

2.1 Distractor Images

A common aspect of picture-password systems is the use of *distractor* images. In early studies confirming the picture superiority effect, memory was tested by asking participants to choose between a previously learned item and a previously unseen item. In an attempt to replicate the excellent memory for pictures found in those studies, some current picture-password systems ask the user to choose between previously **learned** pictures and previously **unseen** pictures. The previously unseen images have been termed *distractor* images [8, 6, 30].

Though this approach can be expected to increase system usability by making it easier for users to select their password items, it has serious problems from a security perspective. Since the distractor images are changed on each authentication attempt, an *attacker* (a person or program that attempts to steal a user's password) can determine items in the user's password by taking the intersection of two screens. In the VIP2 system, distractor images are not changed in case of authentication failure [6]. However, this does not eliminate the problem, as an attacker could simply collect authentication screens between successful authentications and compare them after a period of time.

Further, the use of distractor images makes *hashing* impractical for the system in question. Hashing is a technique employed in almost all character-based password systems that allows user input to be compared to a previously chosen value without storing the value explicitly. Instead, a one-way encryption (the *hash*) of the user's password is stored and compared with the hash of the user's input. If they match, the user has successfully authenticated. The hash function is not easily reversible, so if an attacker gains access to the password file on the server, they cannot simply read user passwords from the file. However, the use of distractor images requires the authentication server to "know" the user's password items, because it must substitute other images in place of non-password items. The authors of previous studies were aware of this, and assumed the existence of a secure authentication server. Though this is often an acceptable assumption, it is suspect in cases involving local authentication. These include logging on to a shared PC, or systems which locally cache server authentication credentials for use when the server is unavailable [14]. In our picture-password system, support for hashing was an important design criteria so distractor images were not used.

2.2 Picture Superiority

The picture-password system described in this paper used a subset of images from the SVLO picture set [16, 17]. Rossion and Pourtois developed the SVLO set by adding color and details to the Snodgrass and Vanderwart picture set [24] which is commonly used in picture memory studies. Though the picture superiority effect has been confirmed several times, the cause of the effect is a subject of continuous debate [11]. However, there are several significant findings which informed the design of the system described here.

Deregowski and Jahoda found that objects are remembered better than pictures which are, in turn, remembered better than text [7]. This supports a model, proposed by Nelson [15], in which pictures are more easily remembered because their semantic meaning is more easily understood. Therefore, pictures from the SVLO set were chosen that most successfully represented objects and were still recognizable when resized. This was accomplished by collecting data about the pictures from three different studies [26, 16, 24] and aggregating the results.

Reversal of the picture superiority effect has occurred in tasks involving similar pictures [23]. Unfortunately, the problem of picture "similarity" is a difficult one, because pictures may be similar along any one of several dimensions: conceptually, verbally, schematically (when two pictures have spatially similar representations), etc. Therefore, the SVLO set was further filtered to remove images with similar verbal labels or schematic similarity.

2.3 Shoulder Surfing

An obvious argument against picture-password systems is that the user's password is easily revealed to an observer if the system's interface uses an on-screen cursor. Several approaches have been developed to deal with this problem but we propose that they be categorized into two groups: *shoulder-surfing resistant* (SSR) and *shoulder-surfing immune* (SSI) authentication systems.

SSR systems, such as the PassFaces^{TM}variant by Tari et al. [28], and the Spy-Resistant Keyboard [27], allow for secure authentication while being observed by a live observer, but are not secure to recording devices. They typically work by obfuscating user input so that by the time an input is observed, it is too late for an attacker to map that input back to the user's password. However, if a shoulder-surfing resistant input is recorded, and then played backwards or forwards multiple times, the password items can be stolen.

SSI systems, such as the Convex Hull Click scheme [33], the cognitive trapdoor system [18], or portfolio-based systems [12], do not attempt to obfuscate user input. Instead, the user is presented with randomized challenges which can only be successfully completed with knowledge of the user's password, but do not directly reveal the password itself. An attacker can only authenticate as the user if they are presented with the same challenge as observed (which is unlikely due to the randomization of the challenge) or by determining the user's password over several successful authentications.

A major distinction between SSR and SSI authentication systems, from a security design perspective, is support for hashing. As with use of distractor images, SSI systems require the authentication server to have explicit knowledge of the user's password and therefore do not support hashing. In a sense, security on the server side is traded for security on the client.

Since support for hashing was an important design factor of our system, an SSR input mode was included. Further, the system was designed to support using the keyboard exclusively for input (though on-screen interaction is still available to users.) This provides at least equal protection against shoulder surfing as a character-based password system.

3 Password System and Experiment Design

In our study, a picture-based password system was compared with a character-based system. Unlike previous studies however, the use of randomly-assigned passwords allows the entropy of passwords in both systems to be known. A comparison of character and picture-password systems with equal entropy has not been previously reported.

The study used a between-subjects design involving two groups. Each participant in the "picture" group received a single picture-based password, while the "character" group received character-based passwords. All passwords were assigned from a set of 80 pictures or characters and consisted of eight items each. Character passwords were composed of eight randomly selected characters from the set shown in Figure 1 and picture passwords were composed of items from the set shown in Figure 2. There were no repeated items within passwords, though no restriction was placed on repeated items between passwords. The entropy of passwords of both types was slightly greater than 50 bits, where each password could take on 1.17×10^{15} possible values.

3.1 Picture-Password System

In our picture-password system, each picture is labeled with a character corresponding to a key on the keyboard. The keys are always

!	@	#	$	%	^	&	*	(	)
Q	W	E	R	T	Y	U	I	O	P
A	S	D	F	G	H	J	K	L	:
Z	X	C	V	B	N	M	<	>	?
1	2	3	4	5	6	7	8	9	0
q	w	e	r	t	y	u	i	o	p
a	s	d	f	g	h	j	k	l	;
z	x	c	v	b	n	m	,	.	/

Figure 1: Character set

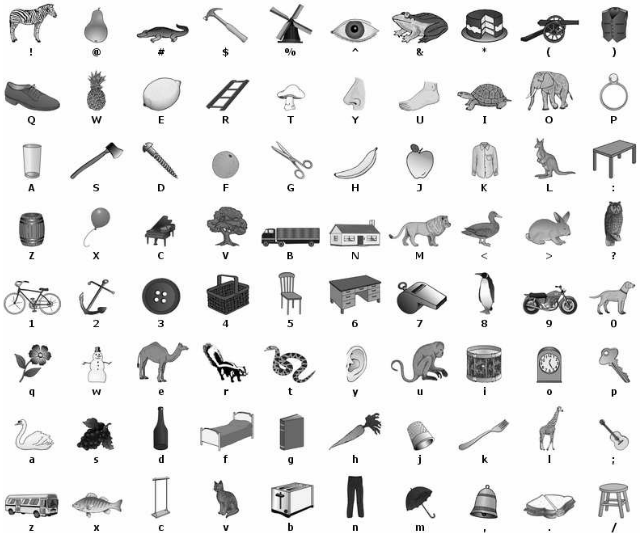


Figure 2: Picture set

arranged in the manner shown in Figure 1 (compare with Figure 2) regardless of the arrangement of the pictures. This corresponds to uppercase and lowercase versions of four rows of keys on a standard, *qwerty*-style keyboard.

During the first stage of training, participants were required to complete one trial exclusively with the mouse, and one with the keyboard. After this stage, participants were free to use the keyboard or an on-screen mouse cursor to enter their password and could mix interaction styles if desired.

Each participant received a unique arrangement of pictures, known as their *home* grid. Participants worked exclusively with their home grid until the final day of testing. In the home grid, pictures are always found in the same location and correspond to the same keyboard key. This enables users to have a consistent input task when entering their password, whether using the keyboard or an on-screen mouse cursor. It was predicted that users would initially use the on-screen mouse cursor to enter their password, and later switch to keyboard input for speed. This behavior would potentially reinforce three redundant encodings of the password: pictorial, verbal, and spatial, which would have an enhancing effect on memory [15].

3.2 Experimental Stages

Participants performed their tasks individually within a nine-day period (though sometimes it was longer than nine days due to scheduling issues.) They first came into the lab on *Day 1*, were tested on *Day 2*, and retested on *Day 9*.

Both character and picture-based participants underwent the same stages of training and testing on Day 1 and Day 2. Additionally, picture-based participants performed the SSR task on Day 9 after memorability data had been collected, and also completed an evaluation survey. The series of tasks performed by participants is given in Table 1. Both password systems were implemented as Java Web Start applications to allow for remotely completing the task on Day 2.

Stages 2, 4 and the SSR task are explained in more detail in later sections. The remaining stages were straightforward password input tasks: picture-password participants selected their password items from a static picture grid, and character group participants entered their password into an empty text box. For all tasks on Day 1, participants were able to restart training if they were unable to recall their password.

Table 1: Experiment tasks

Task Name	Inputs	Description
Day 1 - In Lab		
Stage 1	2	Practice input while password is shown
Stage 2	4	Learn password interactively
Stage 3	4	Enter password with no assistance
Stage 4		Empty screen / consolidation stage
Stage 5	4	Reenter password with no assistance
Day 2 - Any Location		
Day 2	1	Unsupervised entry performed via website
Day 9 - Character Group - In Lab		
Day 9	2	Supervised entry
Day 9 - Picture Group - In Lab		
Day 9	2	Picture Group 1 - Supervised entry performed on home grid, then randomized grid Picture Group 2 - Supervised entry performed on randomized grid, then home grid
SSR Input	1	Shoulder-surfing resistant input task
Evaluation		Evaluation survey of the picture-password system

3.2.1 Day 1 - Stage 2 - Interactive Learning

Previous studies in picture-password systems always involved a training system [31, 8]. There are two reasons for this:

1. Participants usually have not had experience with a picture-based password system before and need guidance in its operation.

2. It is necessary to make sure participants have successfully learned their password by some objective criteria before testing retention over several days.

In previous studies, training was accomplished by requiring the user to perform some number of correct inputs without assistance before continuing. This study followed the same design, requiring eight complete correct inputs, without assistance, by the end of the participants' first day. However, the design of a training system can actively affect the way users learn passwords.[4]

Contemporary theories of serial learning agree that item order is remembered via positional associations rather than through association with previous elements [1, 10]. In other words, in memory, items in lists are associated with their ordinal position in that list. According to Johnson's model of serial learning, association with other password elements may confound retrieval if those elements become associated with other ordinal positions. Because of this, only one item in the user's password is displayed at a time in Stage 2. This is illustrated and explained in Figure 3.

Since the participant focused on their password items while inputting them with the on-screen mouse cursor and the keyboard in Stage 1, they should have little difficulty in finding their pictures in Stage 2 since the other pictures were never explicitly learned.[5] Once the first item is found, subsequent items can also be found interactively by observing the grid and watching for changes.

Various cues to their password are also made visually available to users during this stage:

- The neighboring images to each password item should serve as cues to each item. Based on Johnson's model [10], not including other password items should prevent positional confusion.

- Participants must actively watch the grid while learning their password in order to find their password items. This should reinforce spatial relationships between items.

- In addition to providing visual feedback when items are selected, the asterisks in the password field also associate password items with an ordinal position.

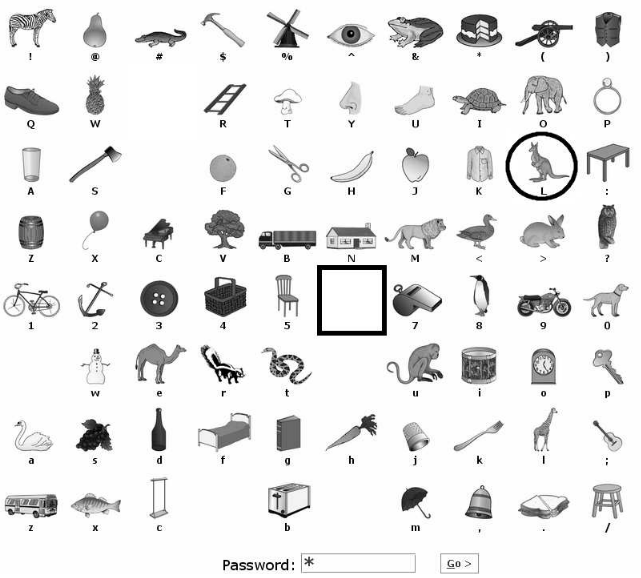
Character-password participants were shown a similar grid, with on-screen characters laid out in the arrangement shown in Figure 1. The operation of this task was the same as that for picture-group participants, with only one item of the user's password shown in the grid at one time.

3.2.2 Day 1 - Stage 4 - Consolidation

In previous studies of picture passwords, it was noted that users spend more time learning a picture password than a character password [8]. This may impact memorability. Recently formed memories are fragile, and the biological process by which these memories are strengthened against forgetting is known as **consolidation**. Neuroscience studies confirm that retroactive interference in the consolidation process is the primary factor in forgetting [34].

Even though previous picture-based password studies equalized the number of correct input trials completed by participants, no attempt was made to equalize total time spent learning the password. Therefore, it is possible that character-based password users are often found to have lower performance for remembering passwords simply because they spend less time learning them and encounter retroactive interference from other sources sooner after learning than picture password users.

In Stage 4, learning times are equalized among participants through use of a non-stimulus. Based on pilot testing, a set time of ten minutes was chosen as an upper bound on the length of time picture-password users would spend learning their password. The participant was shown a nearly empty screen with a cross in the center. Participants viewed this screen until their total time since first seeing their password was ten minutes. Additionally, participants were verbally instructed not to look around the room.



The "kangaroo" (circled) is the participant's current password item. Once selected, the next item (the "desk", see Figure 2) will appear in the space marked by the dark box. The other six items in the password are currently invisible. The circle and box shown here are for illustrative purposes only and did not appear on the participant's screen.

Figure 3: Stage 2 - Picture group

3.2.3 Day 9 - Shoulder-Surfing Resistant Input

As explained in section 2.3, an SSR task was tested with our picture-password system. Participants were shown a randomized grid of pictures, where letters were no longer displayed beneath each picture (see Figure 4). Selection of items by on-screen mouse cursor was not enabled for this task, although the mouse pointer was still visible. Participants were required to hold down a toggle switch, `Ctrl`, to see a grid of keys mapped to pictures in their password (see Figure 5). Participants then needed to press the key which appeared in the same location as their password item (it was not necessary to hold down `Ctrl` at this time.) Upon pressing any key, the characters reshuffled while the picture grid remained unchanged. This pattern of entry continued until the participant's complete password had been entered.

The rationale behind this is similar to the Spy-resistant Keyboard [27] in that screen elements are randomized on each input. It is further strengthened against observation by restricting use of the on-screen cursor.

4 Results and Discussion

23 participants were recruited from students and staff and each was assigned a character or picture-based password based on order of enrollment. 15 participants received picture-based passwords and 8 received character-based passwords.[6] All participants were instructed not to write down their password at any time.

[4] The following discussion of our training system assumes users initially receive their password in a secure location. Note that this is a requirement of any randomly-assigned password system.

[5] This corresponds to a *recognition* task, as the other pictures can be considered *distractor images* at this early stage (see section 2.1).

[6] The unbalanced group sizes were due to an experimental factor discussed in section 4.3. Statistical tests appropriate for unbalanced groups, such as the Fisher Exact Test and Welch's t-test, are used throughout this paper.

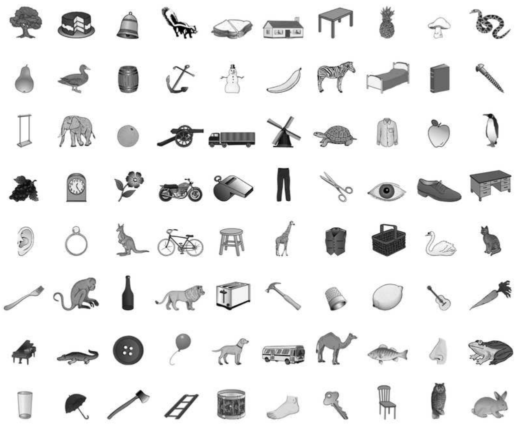
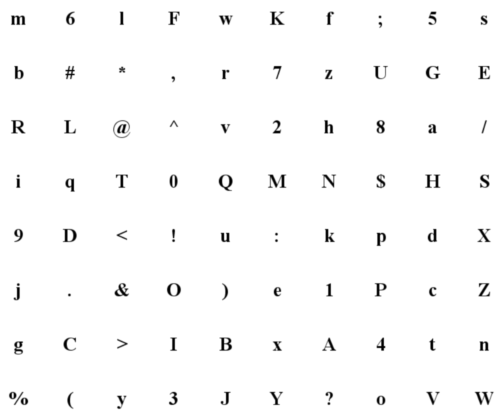


Figure 4: SSR task - Static grid



m	6	l	F	w	K	f	;	5	s
b	#	*	,	r	7	z	U	G	E
R	L	@	^	v	2	h	8	a	/
i	q	T	0	Q	M	N	$	H	S
9	D	<	!	u	:	k	p	d	X
j	.	&	O	)	e	1	P	c	Z
g	C	>	I	B	x	A	4	t	n
%	(	y	3	J	Y	?	o	V	W

Figure 5: SSR task - Ctrl toggled dynamic grid

4.1 Memorability

4.1.1 Ordered vs Unordered Recall

Memorability of a password is measured here based on the participant's ability to enter their password correctly within five tries. It was tested after one day and again after one week. After the experiment concluded, input data for all participants was parsed to determine how successful participants **would have been** at an unordered input task. An *unordered input* involves entering all password items in any order. For example, if the user's password is "EwIg7$cw", "Igcw7Ew$" would be an acceptable unordered input. Similarly, for picture passwords, selecting the correct images in any order is acceptable. Note that participants were never trained on an unordered input task and were not aware that an unordered input would have been accepted as "correct." The analysis of unordered inputs was not performed until after all participants had completed the study and all data had been collected.

4.1.2 Memorability after One Day

Results from the Day 2 task are given in Table 2. The results show that picture passwords were more memorable than character passwords in both ordered and unordered form, though not significantly so. All 15 picture-group participants successfully entered their password in correct serial order. Of character-group participants, 2 out of 8 were unable to correctly enter their password within five tries. However, 1 of the 2 participants made only a transposition error in their password and would have succeeded at an unordered input task.

Fisher's exact tests were run on the 2×2 matrices of Table 2 and the results are given below each table. Mean time between Day 1 and Day 2 across all participants was 38.6 hours, which is much longer than the expected 24 hour interval. This was partially due to webserver downtime for a three day period that pushed back the Day 2 task for five participants (2 in the character-password group and 3 in the picture-password group). Character-password participants had a longer interval between tasks ($\approx$ 45 hours) than picture-password participants ($\approx$ 35 hours) but not significantly so. This was confirmed by Welch's *t*-test (t = 0.73, df = 15.73, p = 0.48).

4.1.3 Memorability after One Week

Results from the Day 9 task are given in Table 3. Character and picture passwords are compared using the *home* grid condition for the picture-password group, and the two character-group participants who failed to authenticate on Day 2 were not brought back for Day 9. The results again show that picture passwords were more memorable than character passwords in both ordered and unordered form, though not significantly so. However, in the unordered analysis, the difference appears to be marginally significant (Fisher's $p \approx 0.07$)[7].

Only 10 of 15 picture-group participants successfully entered their password in correct serial order. Of the character-group participants, three out of six were unable to correctly enter their password within five tries. However, only one of the three character-group

Table 2: Memorability results after one day

Ordered Input

	Correct	Incorrect	% Successful
Character	6	2	75%
Picture	15	0	100%

Fisher's $p = 0.1107$

Unordered Analysis

	Correct	Incorrect	% Successful
Character	7	1	87.5%
Picture	15	0	100%

Fisher's $p = 0.3478$

Table 3: Memorability results after one week

Ordered Input

	Correct	Incorrect	% Successful
Character	3	3	50%
Picture	10	5	67%

Fisher's $p = 0.631$

Unordered Analysis

	Correct	Incorrect	% Successful
Character	4	2	67%
Picture	15	0	100%

Fisher's $p = 0.071$

[7] Removing two participants from the character-password group reduced the already small sample size. If memorability results are extrapolated, with the missing participants favored as to their performance on Day 9, unordered picture passwords appear to be significantly more memorable than unordered character passwords (Fisher's $p < 0.05$).

participants would have succeeded at an unordered input task, as the other two participants entered characters which were not in their password. Among picture-password participants, all 15 chose only the items in their password from their home grid, making only transposition errors with their passwords. More surprisingly, 14 of the 15 picture-group participants chose the correct password items on their first attempt, with the remaining participant requiring only two attempts.

Fisher's exact tests were again run on the 2×2 matrices of Table 3 as shown. Mean time between Day 2 and Day 9 across all participants was 166.1 hours, which is very close to the expected 7 days between Day 2 and Day 9. In this case, picture-password participants had a longer interval between tasks ($\approx$ 168 hours) than character-password participants ($\approx$ 160 hours), but this difference was again not significant (t = -0.99, df = 6.25, p = 0.36).

4.1.4 Discussion

The memorability of picture-password items after one week was 100%. This is striking, but not significantly different from the memorability of character passwords (67%) at this sample size. A randomly-assigned password that consists purely of eight items, with no serial ordering component, has an entropy of $\approx$ 35 bits. This is as secure as a user-selected password of length 12 with strict password policies [4], which should be sufficiently strong for most organizations.

Conversely, the memorability of both picture and character-based passwords, in serial order, was extremely poor. After one week, only 50% and 67% of the character and picture-groups respectively were able to enter their password in five tries. These passwords have an entropy of $\approx$ 50 bits and this result suggests that ordered passwords of this entropy level are too difficult to remember to be practical.

4.2 Comparison of Entry Times

A comparison of entry times was made based on data from Day 2, where participants experienced a short delay ($\approx$ 39 hours) since last entering their password. Two measures were compared: *single* entry time, and *total* entry time. **Single** entry time was taken as the time to enter a single, correct input. This includes corrections, like hitting backspace, but does not include incomplete or incorrect inputs. **Total** entry time was the total time the user spent authenticating, inclusive of failed inputs.

For single entry, the mean times were 10.5 s for characters and 13.7 s for pictures. For total entry, the mean times were 10.5 s and 22.4 s for characters and pictures respectively. The minimum single entry times were 6.6 s and 5.3 s for characters and pictures respectively.

Welch's t-tests were performed for single entry times (t = -1.15, df = 17.96, p = 0.26) and total entry times (t = -1.68, df = 12.93, p = 0.12) which suggest that mean entry times between groups were not significantly different. Due to the small sample size, there may not be sufficient evidence to accept the null hypothesis that the means of both groups are equal. However, the fact that the fastest entry time for all participants was a picture password suggests that variance in picture entry time is large enough that there may be no significant difference between groups.

4.3 Factors within the Picture-Password Group

The two-to-one ratio in group size between picture and character-password participants was chosen due to two independent factors tested within the picture group.

On Day 9, half of the picture-group participants were asked to enter their password on a randomized grid before entering their password on their home grid, while the other half entered their password on their home grid first. Users were not made aware of this task before they encountered it. There was **no significant effect** of viewing the randomized grid on performance in the home grid condition. Because of this, the picture-password group is considered as a whole when compared with the character group for memorability.

4.3.1 Keyboard Usage

During Day 1, Stage 1, half of the picture-group participants[8] were required to use the keyboard exclusively on their first input, and the other half were required to use the mouse. Use of keyboard first, or mouse first, appears to have had **no effect** on whether or not participants used the keyboard on subsequent inputs. Four of 15 participants chose to use the keyboard beyond Stage 1. Of these four, two had used the mouse first, and two had used the keyboard.

Participants did not transition to the keyboard as expected. As stated previously, entry times for picture passwords were not significantly different than those for character passwords. Further, in evaluating the picture-password system, almost all picture-group participants rated a higher level of satisfaction with using the mouse than the keyboard. The exceptions were three keyboard users, who not only preferred to use the keyboard, but rated the picture-password system as "a lot less efficient" than character-based systems.

4.4 Shoulder-Surfing Resistant Input

All picture-group participants completed an SSR input as described in section 3.2.3. The main purpose of this task was to observe an SSR authentication system in operation and allow users to evaluate it. Two important results are given below:

1. In evaluating the SSR input task, 9 out of 15 participants responded that it was **less secure** or **equally secure** to a character-password system.

2. Six out of 15 participants used the mouse while performing the SSR input task. Though on-screen cursor selection of pictures was disabled, the mouse pointer was still active. These participants used the mouse pointer to keep track of their password items while pressing `Ctrl` to see their corresponding grid of keys. Of these six participants, five gave the **less secure** or **equally secure** response mentioned above.

Both of these items confirm the notion that many users have an inadequate understanding of security. The SSR input task is certainly more secure than a character-password system, but participants did not understand this. However, while this might be correctable only with substantial education about security, item #2 may be correctable by better design. This topic is revisited in section 6.3.

5 Repeated Incorrect Inputs

While analyzing the results of the password study, it was observed that users often enter the same incorrect password several times. This occurred during **individual sessions**, and is not a result of aggregating inputs over multiple authentications. 14 of 23 participants made more than one incorrect input in a single session, and of the total number of incorrect inputs made by these users, an average of 29% were repeats from the **same** session.

When a password is incorrectly remembered, users might be quite sure of their password, and unaware of their error. It can be assumed that users repeatedly enter the same incorrect password because they believe it to be correct. They might believe that a typo prevented acceptance of the input. Since users cannot see the actual text entered into a password box, they have no way of knowing if what was submitted to the system was what they had intended.

Regardless of the reasoning, after repeatedly inputting an incorrect string a few times, users will eventually move on to other

[8] These participants were evenly split between the randomized-first and home grid-first groups mentioned above.

guesses, sometimes hitting upon the correct password. Users do not have perfect recall (or perfect typing skills) and allowing multiple guesses at entering a password increases password usability [19]. Unfortunately, if a user makes too many incorrect guesses, they will be *locked out* by many authentication systems.

Lockout may refer to a complete inability to login to an account, or an excessive delay required by the server between authentication attempts. This occurs after a preset number of failed logins have been exceeded. The primary purpose of lockout is to repel attackers who are attempting to guess a user's password. However, when users enter the same incorrect input repeatedly, they use up their available login attempts. This causes the user's account to be locked out very quickly.

Three of the participants in the password study made three repeated incorrect inputs in a single session, though they were aware of the failed-login limit. For these participants, of the five tries available, four consisted of the exact same incorrect string (in two cases, these inputs were non-consecutive.) This resulted in each user being locked out after only making two distinct guesses. Ignoring this inherent user behavior rewards attackers with the ability to make a full complement of guesses, while penalizing legitimate users by reducing the number of distinct guesses they can make.

5.1 Ignoring Duplicates of Incorrect Inputs

Password systems can be expanded to ignore repeated inputs by storing hashes of the inputs in a Temporary Incorrect Input List (TIIL). When a matching input is encountered, it should be handled without penalizing the user. Separate TIILs must be maintained for each user, and the TIIL should be cleared whenever the following conditions are met: a) the user successfully authenticates, b) the account is locked out, or c) the number of login attempts available to the user is reset.

Clearing the TIIL on these conditions restricts the size of the TIIL to the number of login attempts defined by the system. Unless the size of the TIIL is known to the attacker, they will be unable to determine when an extra authentication attempt has been granted because they will never receive more login attempts than the system limit. Even if the attacker is somehow able to determine the size of the TIIL, the risk of using TIILs is very low because the size of the TIIL is inversely related to the number of guesses an attacker can make. Using the clearing policy given above, the TIILs will almost always be empty, so the risk of determining TIIL items through an offline attack is also very low. If TIILs had been implemented in our study, the three participants mentioned above would have received a full five guesses at their password and might have authenticated successfully.

6 Future Work

6.1 Order and Entropy

Both the character and picture passwords in our study were not memorable. We hypothesize that this is primarily due to the requirement of serial ordering, but this hypothesis has not been experimentally verified. It may be that passwords of very high entropy ($\approx$ 50 bits in our study) are simply too difficult to remember. A study involving unordered passwords of this entropy level (containing 14 items for example) would be informative.

6.2 Unordered Passwords

The results of our study suggest that unordered recall could produce an extremely successful authentication system. This result needs to be confirmed with a larger sample size.

Since users were trained serially, it is also necessary to see whether this result is still maintained if users are not trained with a serial requirement. In this case, the design of the training system may have an effect on the performance of participants. Though password items may be entered in any order, users still choose the items in some particular order. Training can then be carried out with three approaches:

1. Impose no order during training. Users are allowed to select items as freely in training as in a typical authentication.

2. Allow users to choose an order during training. Users are asked to select the items in a particular order, and are constrained to that ordering for the duration of training.

3. Generate an ordering for users and constrain them to that ordering for the duration of training.

The intention here is to determine if memorability improves when users learn a particular order during training, even when passwords are accepted without respect to order.

Generating an ordering may be helpful for some users. When the set of password items for a user is randomly generated, the system could find an ordering that enhances memorability and present it to the user. For example, it might choose an ordering of items that alternates (as much as possible) between hands when entered with the keyboard, or an ordering of pictures that runs from left to right and top to bottom on the display, or in a rough circle. Users could then accept this ordering, choose their own, or perhaps have the system generate another ordering based on a different algorithm. There are $n!$ orderings for an unordered password of length n, so having the system generate orderings would likely be beneficial to users.

This does not reduce the strength of the passwords, because the underlying items are still randomly chosen. The unordered passwords still have an entropy of $\approx$ 35 bits. Allowing user choice in ordering has no effect on this.

6.3 Shoulder-Surfing Resistant Input

The SSR input method was not very successful. Six out of 15 participants compromised the security of the system by using the on-screen mouse cursor to keep track of password items. Though simply hiding the on-screen cursor may seem like a solution to the problem, it is evident from the participants' behavior that keeping track of an unfamiliar location in a 10x8 grid is too difficult for users. Even if the mouse pointer were made invisible, the difficulty of the task may force users to use their finger to keep track of password items. Framing the grid with lines or row/column identifiers might make it easier to keep track of item locations.

6.4 Long-Term Effects

The majority of participants in the picture-password group used the mouse instead of the keyboard. It was predicted that, in the interest of speed, users would eventually settle into keyboard entry. A longer term study, with more frequent input tasks, is needed to determine whether or not this is true. If true, determining the effect this has on picture-password memorability, and the reconstruction of partially forgotten passwords, is also important.

7 Conclusions

Across all conditions, picture passwords were more memorable than character passwords, though the difference was not significant due to the small sample size of the study. It was marginally significant when input data was analyzed to determine how well participants would have performed at an unordered input task. In this case, the picture-password group performed better than the character group after one week: 100% recall and 67% respectively, with all 15 picture-group participants correctly selecting only their password items within two tries. This appears to be a confirmation of the picture superiority effect, and may also be attributable to the "multiple encodings" of each password item (each item was represented by a picture as well as a keyboard key and location in the home grid.)

However, when ordered passwords with a full 50 bits of entropy were considered, performance for both picture and character passwords was quite poor: 67% recall and 50% respectively. Serial order information either does not benefit from the multiple encodings of picture-password items or passwords at this entropy level are too difficult to remember.

A couple observations about user behavior were also made. Most importantly, the fact that users repeat incorrect inputs is likely based on the fact that users do not receive adequate feedback when entering a password (they cannot see the actual text submitted to the system.) Since this is unavoidable for security reasons, the duplicate inputs should be discarded by the authentication server and not counted against the user. This does not compromise the security of the system, since attackers have nothing to gain from duplicating inputs.

User behavior during the SSR task was unexpected. The purpose of the task had been explained immediately before it was performed, yet six out of 15 participants revealed their password through an insecure behavior. This highlights the importance of usability testing in security applications.

Picture passwords are a relatively new area of study, so the possibilities for future work are extensive. Based on the results presented here, the most promising future work is in the area of unordered, randomly-assigned passwords. Research into insecure user behaviors and training methods is also extremely important.

References

[1] J. Anderson and M. Matessa. A production system theory of serial memory. *Psychological Review*, 104(4):728–748, 1997.

[2] G. Blonder. Graphical password, Sept. 24 1996. US Patent 5,559,961.

[3] S. Brostoff and M. Sasse. Are Passfaces more usable than passwords? A field trial investigation. In *People and Computers XIV-Usability or Else!*, pages 405–424, 2000.

[4] W. E. Burr, D. F. Dodson, and W. T. Polk. Nist special publication 800-63. *Electronic Authentication Guideline,? Version*, 1, 2004.

[5] D. Davis, F. Monrose, and M. Reiter. On User Choice in Graphical Password Schemes. In *13th USENIX Security Symposium*, pages 151–164, 2004.

[6] 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.

[7] J. Deregowski and G. Jahoda. Efficacy of Objects, Pictures and Words in a Simple Learning Task. *International Journal of Psychology*, 10(1):19–25, 1975.

[8] R. Dhamija and A. Perrig. Deja Vu: A User Study Using Images for Authentication. In *Proceedings of the 9th USENIX Security Symposium.*, pages 45–48, 2000.

[9] A. Dirik, N. Memon, and J. Birget. Modeling user choice in the PassPoints graphical password scheme. *Proceedings of the 3rd Symposium on Usable Privacy and Security*, pages 20–28, 2007.

[10] G. Johnson. A distinctiveness model of serial learning. *Psychological Review*, 98(2):204–217, 1991.

[11] H. Kinjo and J. Snodgrass. Is there a picture superiority effect in perceptual implicit tasks? *European Journal of Cognitive Psychology*, 12(2):145–164, 2000.

[12] S. Man, D. Hong, and M. Mathews. A shoulder-surfing resistant graphical password scheme. In *Proceedings of International conference on security and management*, volume I, pages 101–111, 2003.

[13] J. Massey. Guessing and entropy. In *Proceedings of the IEEE International Symposium on Information Theory*, 1994.

[14] Microsoft Corporation. Cached domain logon information. http://support.microsoft.com/kb/172931 (accessed October 2007), 2007.

[15] D. Nelson. Learning to Order Pictures and Words: A Model of Sensory and Semantic Encoding. *Journal of Experimental Psychology: Human Learning and Memory*, 3(5):485–497, 1977.

[16] B. Rossion and G. Pourtois. Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, 33(2):217–236, 2004.

[17] B. Rossion and G. Pourtois. Snodgrass and Vanderwart Like Objects. http://alpha.cog.brown.edu:8200/stimuli/objects/svlo.zip/view (accessed Sept. 2007), 2004.

[18] V. Roth, K. Richter, and R. Freidinger. A PIN-entry method resilient against shoulder surfing. In *Proceedings of the 11th ACM conference on Computer and communications security*, pages 236–245. ACM Press New York, NY, USA, 2004.

[19] M. 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, 2001.

[20] B. Schneier. Schneier on security: Real-world passwords. http://www.schneier.com/blog/archives/2006/12/realworld_passw.html (accessed December 2007), 14 Dec. 2006.

[21] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 1948.

[22] R. Shepard. Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, 6(1):156–163, 1967.

[23] J. Snodgrass and B. McCullough. The role of visual similarity in picture categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12(1):147–154, 1986.

[24] J. Snodgrass and M. Vanderwart. A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 6(2):174–215, 1980.

[25] L. Standing. Learning 10000 pictures. *The Quarterly Journal of Experimental Psychology*, 25(2):207–222, 1973.

[26] G. Stenberg, K. Radeborg, and L. Hedman. The picture superiority effect in a cross-modality recognition task. *Memory and Cognition*, 23(4):425–441, 1995.

[27] D. Tan, P. Keyani, and M. Czerwinski. Spy-resistant keyboard: more secure password entry on public touch screen displays. In *Proceedings of the 19th conference of the computer-human interaction special interest group (CHISIG) of Australia on Computer-human interaction*, pages 1–10. Computer-Human Interaction Special Interest Group (CHISIG) of Australia Narrabundah, Australia, Australia, 2005.

[28] F. Tari, A. Ozok, and S. Holden. A comparison of perceived and real shoulder-surfing risks between alphanumeric and graphical passwords. In *Proceedings of the second symposium on Usable privacy and security*, pages 56–66. ACM Press New York, NY, USA, 2006.

[29] J. Thorpe and P. van Oorschot. Human-Seeded Attacks and Exploiting Hot-Spots in Graphical Passwords. *Proceedings of the 16th Usenix Security Symposium*, pages 103–118, 2007.

[30] D. Weinshall and S. Kirkpatrick. Passwords you'll never forget, but can't recall. In *Conference on Human Factors in Computing Systems*, pages 1399–1402. ACM Press New York, NY, USA, 2004.

[31] S. Wiedenbeck, J. Waters, J. Birget, A. Brodskiy, and N. Memon. Authentication using graphical passwords: Basic results. In *Human-Computer Interaction International 2005*, 2005.

[32] S. Wiedenbeck, J. Waters, J. Birget, A. Brodskiy, and N. Memon. Authentication using graphical passwords: effects of tolerance and image choice. In *Proceedings of the 2005 symposium on Usable privacy and security*, pages 1–12. ACM Press New York, NY, USA, 2005.

[33] S. Wiedenbeck, J. Waters, L. Sobrado, and J. Birget. Design and evaluation of a shoulder-surfing resistant graphical password scheme. In *Proceedings of the working conference on Advanced visual interfaces*, pages 177–184. ACM Press New York, NY, USA, 2006.

[34] J. Wixted. The psychology and neuroscience of forgetting. *Annual Review of Psychology*, 55:235–269, 2004.