A Comprehensive Study of Frequency, Interference, and Training of Multiple Graphical Passwords

Katherine M. Everitt, Tanya Bragin, James Fogarty, Tadayoshi Kohno
Computer Science & Engineering
DUB Group, University of Washington
Seattle, WA 98195
{everitt, tbragin, jfogarty, yoshi}@cs.washington.edu

ABSTRACT
Graphical password systems have received significant attention as one potential solution to the need for more usable authentication, but nearly all prior work makes the unrealistic assumption of studying a *single* password. This paper presents the first study of *multiple* graphical passwords to systematically examine *frequency* of access to a graphical password, *interference* resulting from interleaving access to multiple graphical passwords, and patterns of access while *training* multiple graphical passwords. We find that all of these factors significantly impact the ease of authenticating using multiple facial graphical passwords. For example, participants who accessed four different graphical passwords per week were ten times more likely to completely fail to authenticate than participants who accessed a single password once per week. Our results underscore the need for more realistic evaluations of the use of *multiple* graphical passwords, have a number of implications for the adoption of graphical password systems, and provide a new basis for comparing proposed graphical password systems.

Author Keywords
Graphical passwords, usable security, authentication.

ACM Classification Keywords
H5.2. Information interfaces and presentation. K.6.5 Management of Computing and Information Systems: Security and Protection.

INTRODUCTION
Most people find it difficult to remember alphanumeric passwords [10, 12], a problem magnified by the fact that an average Web user has passwords on 25 unique Web sites [12]. This difficulty leads people to adopt a number of unsafe strategies, including writing passwords down, reusing the same password, using minor variants of a single password, or frequently reinitializing passwords upon failure to authenticate [1, 2, 3, 13, 14, 16]. All of these behaviors increase the likelihood of passwords being lost, stolen, or compromised.

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.

Graphical password systems have received significant attention as one potential solution to the need for more usable authentication [4, 5, 6, 7, 8, 15, 20, 23]. Graphical password systems take many forms, such as requiring the selection of target images from sets of distracter images [8, 17] or requiring clicks on target regions of an image [23]. Graphical passwords are generally considered to be easier to remember and use than alphanumeric passwords because graphical passwords take the proven approach of relying upon *recognition* instead of requiring *recall* [18]. A separate advantage of graphical passwords is their natural appropriateness for situations where text entry is difficult or limited (e.g., when using a small mobile device with limited keyboard input, such as popular touchscreen phones).

Some graphical password systems provide a level of strength (entropy) against password guessing attacks that is equal to or greater than typical alphanumeric passwords, but this is not a strict requirement. Instead, it is clear that different approaches exist at a range of points in a trade-off between usability and cryptographic strength. When password usability is important to an application, even a weak (low entropy) password system can provide sufficient security when used as part of a larger multi-factor authentication system. Widely used four-digit PINs, for example, are typically paired with the need to physically possess an ATM card and a limit on the number of failed attempts allowed before the ATM card is confiscated.

The continuing emergence of the mobile Web seems to promise many additional opportunities for multi-factor approaches. A social networking site, for example, may want to reduce the burden of mobile authentication, but mobile text entry is relatively difficult (especially for the special characters and non-word sequences common in passwords). The site might therefore require that a device initially be authenticated using an alphanumeric password, but then place a cookie on the device. Future access could then use a combination of the cookie on the authenticated device and an easier graphical password. As in the ATM card example, this cookie could be revoked after as little as a single failed attempt at the graphical password. This system would allow people to easily access protected sites from their mobile devices, but even the use of a weak password will guard against illegal access to those sites by someone who might have found or stolen the device.

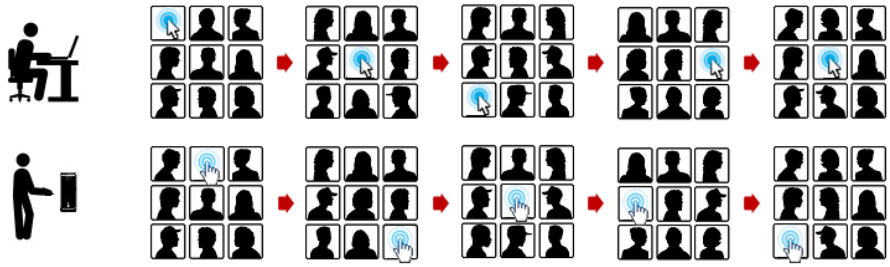


Figure 1: A person attempting to authenticate with a facial graphical password is presented a sequence of 3x3 grids of faces. Successful authentication requires choosing the correct face from each set. We selected facial graphical passwords for study in this work because of their commercial deployment in the PassFaces™ system [17] and because of their use in prior research [4, 9, 21]. Such a system is appropriate for many situations, including the desktop and in mobile situations where text input is more difficult.

Given the need for more usable authentication and existing interest in graphical passwords as a potential solution, we identify an important limitation of existing work: although there have been many studies of graphical passwords, nearly all prior work focuses on a *single* password. People will need to remember and use many graphical passwords, just as they currently use many alphanumeric passwords, but no work has systematically explored the use of multiple graphical passwords.

This paper therefore studies 100 participants using multiple graphical passwords for more than a month. We study *facial* graphical passwords, selected because of their commercial deployment in the PassFaces™ system [17] and because of their use in prior research [4, 9, 21, 22]. Figure 1 illustrates the authentication process, which is based on the presentation of a sequence of 3x3 grids of faces. Successful authentication requires that a person select the correct face from each grid in the sequence, and the length of the sequence can vary according to the needs of the application. Such a system is appropriate for authentication in many environments, including the traditional desktop and in mobile situations where text input is more difficult.

We developed an email-based methodology for studying participant use of multiple facial graphical passwords, wherein participants agreed to receive one to four emails per week. Each email provided a link to our study website and prompted the participant to visit and authenticate. Participants thus accessed our facial graphical passwords in their normal environments from which they might access any other service. By varying the frequency of email to each participant, how many passwords were issued to each participant, and the pattern of access for each participant, we systematically study the use of multiple graphical passwords. Our study is the first of its kind, and our results demonstrate a clear need for future evaluations of graphical passwords to consider the more realistic demands of using multiple graphical passwords.

The contributions of this paper include:

- We *identify* the need to systematically study the use of multiple graphical passwords, as nearly all prior studies of the memorability of graphical passwords are focused on the unrealistic use of a single password.
- We show that *frequency* of access to a facial graphical password significantly impacts ease of access. Participants who accessed a facial graphical password once per week required more attempts and more time to authenticate than participants who accessed a facial graphical password once per day.
- We show that *interference* from interleaving access to multiple facial graphical passwords significantly impacts ease of access. Our findings include the fact that participants accessing four facial graphical passwords per week were ten times more likely to completely fail to authenticate than participants accessing a single facial graphical password once per week.
- We show that patterns of access while *training* multiple facial graphical passwords significantly impact later ease of access. Participants who trained on multiple graphical passwords each week for a month were four times more likely to later completely fail to authenticate than participants who focused on their training a single password during each week of the month.
- We show that *long-term recall* of facial graphical passwords is significantly impacted by *interference* and *training*. Participants who had used only one facial graphical password were still able to access that password four months after completing our study, but the success of participants who had used four passwords was strongly influenced by their pattern of access during *training*.
- We discuss the *implications* of our findings. Our findings regarding *frequency* and *interference* show that field studies of graphical password systems are likely to overestimate ease of access if they do not study the realistic use of multiple graphical passwords. Our findings regarding *interference*, *training*, and *long-term recall* also motivate a need for future field studies examining how people typically acquire and learn new passwords.
- We contribute a study *methodology* that enables realistic and practical studies of the use of multiple graphical passwords. Evidence of our method's effectiveness can be seen in the results we present throughout this paper and in our high participation rate relative to prior work.

RELATED WORK

Existing interest in graphical passwords is motivated in part by results from human psychology research. One such result is that human ability for *recognition* far exceeds that for *recall*. Rock and Engelstein, for example, found that participants shown a meaningless shape were unable to draw the shape a month later (which would require recall), but could recognize the shape almost perfectly [18]. Another relevant result is that human memory for images far exceeds that for verbal material. Standing shows that people can easily recognize thousands of images, that the superiority of images persists even in the context of large numbers of distracters, and that the ability to recognize images persists over time [19]. A full survey of potentially relevant psychology literature is beyond the scope of this section, as is a complete discussion of other approaches to leveraging recognition in usable security (such as SiteKey systems that aim to prevent phishing attacks by presenting an secret image that allows a person to identify a website). We instead focus on graphical password systems, studies of facial graphical passwords, and the limited prior work examining multiple graphical passwords.

Graphical Password Systems

De Angeli *et al.* propose three categories for graphical password systems: *cognometric* schemes, *locimetric* schemes, and *drawmetric* schemes [7].

Cognometric Schemes

Cognometric schemes present a set of images, with authentication requiring selection of the correct images from the set. Images might, for example, be random art images, pictures of faces, or photographs of scenes. Different schemes vary in their selection requirements, such as whether images must be selected in a particular order.

The facial graphical passwords we study in this paper are a cognometric scheme, requiring selection of the correct face from a 3x3 grid of distracter faces. Multiple sets of faces are presented in sequence, with authentication requiring selection of the correct face from each set. We study passwords that use sequences of length five, but sequences of any reasonable length could be used.

Déjà vu uses a set of automatically synthesized random art images [8]. Authentication requires selecting five portfolio images from a set of twenty-five challenge images. In contrast to presenting sequences of sets of images, Déjà vu presents all twenty-five images at the same time. Authentication requires selecting the correct images, *in the correct order*, from the distracters. In a study comparing ease of access using synthesized random art images versus natural photographs, Dhamija and Perrig found a 10% authentication failure rate after one week for the art images versus a 5% failure rate for the natural photographs [8].

Davis *et al.* propose a *Story* scheme, encouraging participants to make up a story to help remember in what order to select a sequence of password images from a set of distracters [6]. They report that participants were often able to remember the correct images, but not the sequence in which they should be selected. They attribute this at least in part to the fact that many of their participants did not actually make up stories, but instead attempted to simply memorize the sequence, and suggest that graphical password systems should avoid relying upon recall of order.

Locimetric Schemes

Locimetric schemes present a single image, with authentication requiring clicking on regions of the image corresponding to a password. One example of such a system is PassPoints [23]. A potential weakness of this class of schemes is the danger of using highly visually salient points that are "obvious" to an attacker [23].

Drawmetric Schemes

Drawmetric schemes require drawing figures or doodles to authenticate. De Angeli *et al.* find that such schemes can be problematic, both because it is difficult to provide context to enable leveraging of recognition (as opposed to recall) and because people may have difficulty recreating the drawing accurately enough to be approved by a system [7].

Studies of Facial Graphical Passwords

Early studies of facial graphical passwords include work by Valentine [21, 22]. This work shows that people are good at recognizing faces and can remember the faces from a single password for months after the initial training. Brostoff and Sasse validate these findings in field situations [4].

Davis *et al.* show that allowing people to select the faces that make up their password can lead to biases toward more attractive faces and toward female faces, significantly reducing security [6]. They find that an attacker can guess 10% of chosen passwords within two tries, and 25% of chosen passwords within thirteen tries (eight tries if the password faces were chosen by a male). Because of this, our study assigns faces to participants. Other work has examined concerns regarding shoulder-surfing [20] and attempts to verbally describe facial graphical passwords [9].

Studies of Multiple Graphical Passwords

We are aware of only two pieces of prior work that include multiple graphical passwords, and neither of these systematically examines the effects of password frequency, interference, or training. Moncur and Leplâtre compare the memorability of five picture-based PINs with five text-based PINs [15]. Their results re-affirm the advantages of graphical passwords, but all participants used the same number of graphical passwords in the same manner, and so no insight is provided into the effects of frequency, interference, or training. Chiasson *et al.* found a difference in authentication success in the extreme situation where two different locimetric passwords are set in *the same image* [5], but do not focus on a systematic examination of frequency, interference, or training. Given the community's significant interest in graphical password systems, it is clear that additional work is needed to further our understanding of the effects of using multiple graphical passwords.

METHOD

Our primary goal was to study the effects of the *frequency* of facial graphical password usage, the effects of *interference* resulting from the use of multiple facial graphical passwords, and the effects of different patterns of access when *training* multiple graphical passwords. We therefore designed a four stage study: (1) a pre-study questionnaire examining participant demographics and current password strategies, (2) a five-week online study of participants accessing multiple facial graphical passwords, (3) a post-study questionnaire regarding participant experiences, and (4) a test of long-term recall conducted four months after the end of the original five-week study.

Design and Procedure

The core of our design is a five-week online study using email-based prompts to access the study website and authenticate. The choice of an online study instead of a laboratory study is a trade-off, but we felt the online study would provide access to a larger participant pool, would mean people authenticated under more realistic settings, and would likely result in a higher retention rate than requiring people make multiple visits to a laboratory. Participants received email in the morning, and the email contained a link to our study website. Study emails were sent on Tuesday, Wednesday, Thursday, and Friday of each week, and the link in each email expired at midnight on the day it was sent. Participants thus accessed our online study at most once per day and at most four times per week.

Authenticating

Upon accessing our online study via an emailed link, participants were presented with the facial password system shown in Figure 2. As we have illustrated in Figure 1, authentication requires selecting the correct face from a sequence of 3x3 grids of distracter faces (we use a sequence of length five in this study). Note that the same sets of faces are presented whenever attempting to authenticate for a particular password, a requirement for preventing attacks based in determining which faces consistently appear across multiple login attempts. Faces are, however, presented at random locations in the 3x3 grid, ensuring that participants need to learn the faces and cannot rely upon spatial positioning. Participants were given feedback only after selecting a face from all five sets, and were allowed a maximum of three login attempts. If unable to successfully authenticate within three attempts, participants were asked to retrain (analogous to resetting the password). Because the facial graphical passwords in our study were not protecting sensitive information, retraining did not change a password but instead reminded the participant of the password and ensured they could authenticate using it.

Study Conditions

We used a between-subjects design with five conditions. Figure 3 summarizes the conditions. We defer extensive discussion of these conditions until our results section in order to discuss the meaningful contrasts between them. Each letter signifies a need to authenticate using a different password. Participants in the first condition, for example, authenticated only once per week during our five-week study, always using the same password. Participants in the second condition authenticated three times per week, again always using the same password. The third condition combines the first two, requiring authentication four times per week (once each week for the first password and three times each week for the second password). The fourth requires a single authentication for each of four different passwords each week, while the final condition initially groups the use of a password within a single week and then requires authentication using all four passwords in the final week. Note that the arrangement of days within each week in Figure 3 is for illustration only, as each week's requirements were assigned to random days. In the first condition, for example, a random day each week was chosen to prompt the participant to authenticate. Similarly, participants in the fourth condition were prompted to access their four passwords in a random order each week.

Figure 2: An example screen from our online facial graphical password system. Participants authenticate by selecting the correct faces from 3x3 grids of distracters.

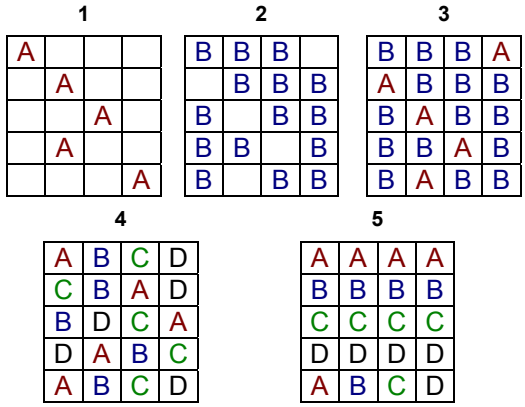


Figure 3: An overview of our five study conditions. Each row represents a week of the study. Each letter represents a different password. Conditions vary in their number of passwords and their patterns of password access. Extensive discussion is deferred to our results sections.

Each password was associated with a mock website that had a distinct logo, name, and background color. This was to ensure participants had a context to use in differentiating their multiple passwords. We also ensured that no face appeared more than once in any of a participant's passwords or distracter faces, as discussed next.

Password Assignment and Training

Passwords were automatically assigned, as prior work has shown that allowing selection of faces results in a bias toward more attractive faces and toward female faces, significantly reducing security [6]. We use faces from the Face of Tomorrow dataset [11], which contains ordinary people (neither models nor celebrities) from a variety of ethnic backgrounds, including a mix of men and women. We manually cropped the images from this dataset to show only the face. This cropping is important because an intentional characteristic of facial graphical passwords is the absence of cues that stand out (background, eye glasses, clothing), thus making it difficult to write down or tell another person a password [9]. Our password assignment ensured there was no overlap of faces appearing in any of a participant's multiple passwords nor their distracter faces. This was done by selecting faces randomly without replacement for each participant. Our results may therefore underestimate interference among multiple services that draw upon the same database of faces (such as a default database shipped with a facial graphical password system).

Participants were shown their assigned password and trained in its use upon their initial attempt to authenticate for each of their multiple passwords (training thus occurred in the first week for all except our last condition). Training consisted of showing a participant their assigned password, asking the participant to briefly think about each face, and asking the participant to authenticate using a version of the interface that highlighted the correct faces. Training was complete when the participant could authenticate using their password without the assistive highlighting.

Measures

We evaluate ease of facial graphical password access using four complimentary measures: (1) authentication *failure rate*, (2) number of *attempts required* for successful authentication, (3) the *login time* required in a successful authentication, and (4) the *total time* required, including time spent on failed attempts.

Our *failure rate* measure examines complete failure to authenticate within the three permitted attempts. Facial graphical passwords are highly memorable [4, 22], and so we expected failure rate to be our least sensitive measure. Whether or not a participant successfully authenticated is a binary measure, so we use chi-squared tests throughout the coming sections when we analyze *failure rate*.

Our second measure examines the *attempts required* before a participant successfully authenticated. This is based in the intuition that a participant who requires all three permitted attempts before successfully authenticating would seem to be having more difficulty than a participant who is able to successfully authenticate on the first attempt. Our analyses find that *attempts required* is not normally distributed (most participants authenticate in a single try), so we analyze *attempts required* using the Mann-Whitney test (reporting the chi-squared approximation of p, consistent with common statistical practices). The Mann-Whitney test is a non-parametric test based in ranking observations across conditions and then testing for differences in rank. In cases where participants failed to authenticate within three attempts, we coded a value of 4 for *attempts required* (note that the choice of 4 is arbitrary, as our use of a rank-based statistical test means that identical results would be obtained using any consistent value greater than 3).

Our final two measures examine the time to authenticate, consistent with prior work that has examined time as a measure of the difficulty of facial graphical password use [21, 22]. In order to separate these measures from *failure rate*, we exclude cases where a participant failed to authenticate within three attempts. The *login time* measure then considers the amount of time spent during a participant's successful attempt (ignoring time spent on unsuccessful attempts). The *total time* measure considers the total time spent authenticating, including time spent on unsuccessful attempts. Because our analyses find that *login time* and *total time* are both non-normally distributed, we again use the Mann-Whitney test throughout our analyses.

Our use of a rank-based statistical test also provides robustness against many types of noise that might affect time measures. For example, our use of an online study means we cannot guarantee participants were not interrupted in the midst of authenticating. Although we see no evidence of abnormally long delays and such a concern seems equally likely in any study condition, it is worth noting that a non-parametric test is extremely resilient to such outliers (for much the same reason that a median is more resistant to outliers than a mean). Similarly, our timing data includes the time needed to load the study webpage, but page load times were minor compared to login times and a non-parametric test will be unaffected.

Participants

We primarily recruited participants from the undergraduate and graduate students in our university's Asian Languages, Electrical Engineering, Computer Science & Engineering, and Material Sciences departments. Participants were recruited via opt-in email and screened to ensure that they had no prior experience with facial graphical passwords or any other graphical password system. Participants were compensated with a $10 iTunes gift certificate.

A total of 110 people agreed to participate, 34 female and 76 male. Our demographic was mostly university students, and 69% of participants were ages 18 to 24, 18% were ages 25 to 29, and 12% were age 30 or older. Participants were randomly assigned to conditions using a round-robin strategy to ensure balanced group size.

Our participation rate was quite good, which we attribute primarily to our email-based methodology and the fact that each day's session was very simple and could be completed quickly. Of the 110 participants who originally agreed to participate, seven had low participation rates because they either never accessed the study website for training (at least one accidentally provided an incorrect email address) or they completed initial training and then rarely or never returned. Another three participants accidentally viewed faces from our site in a manner inconsistent with the study (two due to a minor bug at the beginning of the study and one by accidentally clicking on an email while using a friend's computer, thus visiting the friend's study page). The remainder of this paper therefore focuses on data from 100 participants. These participants responded to 92% of email-based prompts, and 60% of participants responded to every email-based prompt. As a point of comparison, we note that Moncur and Leplatre report a 35% completion rate in a study of multiple graphical passwords that required visiting a Web page three times in four weeks [15].

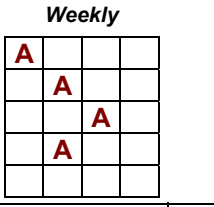
PRE-STUDY QUESTIONNAIRE RESULTS

Participants reported password management experiences that are typical of university students, affirming that password management is generally problematic. Of our 100 participants, 79% reported forgetting a password, 98% reported reusing a password across multiple websites, 87% reported using simple variants of the same password, 65% reported using Web browser support for storing and automatically entering passwords, 51% reported writing down a password, and 52% reported using personal information, such as a birthday, in a password. Only 11% reported trying dedicated secure password storage software. When asked about their number of unique passwords versus the number of sites on which they had passwords, 90% reported having fewer than 10 unique passwords but 78% reported having passwords on more than 10 sites. Some passwords must therefore be reused, indicating the ideal of a unique strong password for each site is far from reality.

FREQUENCY RESULTS

Our first planned analysis examines the *frequency* of use of a facial graphical password. Figure 4 illustrates our planned contrast between *daily* and *weekly* use. We intentionally analyze only the first four weeks of data in our *weekly* condition, matching the four study days within a single week in the *daily* condition. Although studies of graphical passwords generally consider only a single frequency, we hypothesized that *weekly* access would result in more failures, more required attempts, more time spent on a successful authentication, and more total time spent.

A chi-squared test of *failure rate* shows no significant difference ($\chi^2(1, 113) = 1.6$, $p = 0.206$). The distribution of *attempts required* is non-normal, and a Mann-Whitney test finds that participants in the *weekly* condition required significantly more attempts ($\chi^2(1,113) = 5.5$, $p = 0.019$). After excluding failures from our temporal analyses, both *login time* and *total time* are non-normally distributed.



	Weekly		Daily
Failure Rate	1.96%	$\chi^2(1,113) = 1.6$, $p = 0.206$	0%
Attempts Required	1.24 tries	**$\chi^2(1,113) = 5.5$, $p = 0.019$**	1.03 tries
Login Time	18.14 sec	$\chi^2(1,112) = 3.3$, $p = 0.067$	15.56 sec
Total Time	23.59 sec	**$\chi^2(1,112) = 4.6$, $p = 0.031$**	18.25 sec

Figure 4: We examined *frequency* of facial graphical password use by comparing *daily* and *weekly* use. Participants in the *weekly* condition required more login attempts, more time per successful login, and more total authentication time.

Mann-Whitney tests show that participants in the *weekly* condition required marginally more *login time* ($\chi^2(1, 112) = 3.3$, $p = 0.067$) and significantly more *total time* ($\chi^2(1, 112) = 4.6$, $p = 0.031$).

INTERFERENCE RESULTS

Our second planned analysis examines the *interference* resulting from using *multiple* facial graphical passwords. We planned several comparisons examining interference for *frequent* and *infrequent* passwords. For the purposes of our studies, we define a *frequent* password as being accessed three times per study week. We define an *infrequent* password as being accessed once per study week.

Interfering with a Frequent Password

Figure 5 illustrates our planned contrast between our *single frequent* condition and our *infrequent distracter* condition. Note that we here analyze only data from the frequent password (password B in Figure 5). Data from the infrequent distracter (password A in Figure 5) is not included in these analyses because we are focused on how the presence of the infrequent distracter affects the use of the frequent password. We hypothesized that the *infrequent distracter* condition would result in more failures, more required attempts, more time spent on a successful authentication, and more total time spent.

A chi-squared test of *failure rate* shows that the *infrequent distracter* condition resulted in marginally more failures ($\chi^2(1,540) = 3.3$, $p = 0.069$). The distribution of *attempts required* is non-normal, and a Mann-Whitney test finds that participants in the *infrequent distracter* condition required significantly more attempts ($\chi^2(1,540) = 4.0$, $p = 0.044$). After excluding failures from our temporal analyses, both *login time* and *total time* are non-normally distributed. Mann-Whitney tests show that participants in the *infrequent distracter* condition required significantly more *login time* ($\chi^2(1,540) = 16.5$, $p < 0.001$) and significantly more *total time* ($\chi^2(1,540) = 17.1$, $p < 0.001$).

Interfering with an Infrequent Password

Figure 6 illustrates our planned contrast between three conditions examining interference with an infrequent password. The *single infrequent* condition provides a baseline against which we compare a *frequent distracter* and *multiple infrequent distracters*. Note that our *frequent distracter* data is the same data from our *infrequent distracter* condition in the previous subsection, but we have reversed the roles of the two passwords to support the comparisons we make here (excluding data for password B of the *frequent distracter* condition from our current analyses). Neither password was presented to participants as being more or less important than the other. Similarly, all of the passwords in the *multiple infrequent distracters* condition were presented as equally important, and so all of them serve as distracters for each other. This subsection's analyses are therefore based on data for all four infrequent passwords in the *multiple infrequent distracters* condition. We hypothesized that our *single infrequent* condition would be the easiest of the three and that the *multiple infrequent distracters* would be the most difficult, as indicated by all four of our measures.

We analyze *failure rate* using chi-squared tests. These show that *multiple infrequent distracters* resulted in significantly more failures than the *single infrequent* condition ($\chi^2(1,312) = 13.4$, $p < 0.001$). A *frequent distracter* caused marginally more failures than a *single infrequent* password with no distracters ($\chi^2(1,144) = 3.7$, $p = 0.054$), and *multiple infrequent distracters* caused marginally more failures than a single *frequent distracter* ($\chi^2(1,318) = 2.8$, $p = 0.093$). The distribution of *attempts required* is non-normal, and we analyze *attempts required* using Mann-Whitney tests. These show that the *single infrequent* condition required significantly fewer attempts than both the single *frequent distracter* condition ($\chi^2(1,144) = 4.6$, $p = 0.032$) and the *multiple infrequent distracters* condition ($\chi^2(1,312) = 10.7$, $p = 0.001$). After excluding failures from our temporal analyses, we find that both *login time* and *total time* are non-normally distributed. Mann-Whitney tests show differences in *login time* are significant across all three comparisons ($\chi^2(1,137) = 7.9$, $p = 0.005$, $\chi^2(1,274) = 41.8$, $p < 0.001$, $\chi^2(1,275) = 8.8$, $p = 0.003$), as are differences in *total time* ($\chi^2(1,137) = 8.4$, $p = 0.004$, $\chi^2(1,274) = 42.3$, $p < 0.001$, $\chi^2(1,137) = 7.2$, $p = 0.008$).

TRAINING RESULTS

Given the expected difficulty of *multiple infrequent distracters*, we also planned to examine how patterns of password use during *training* affects later ease of access to *multiple* facial graphical passwords. Figure 7 illustrates this contrast. Both conditions examine four passwords. In the *mixed* condition, these passwords are trained in parallel, with each of the four being used once per week during the first four weeks. In the *grouped* condition, the four passwords are trained in series, as each of the first four weeks focuses on a single password. Our analyses in this section exclude data from these first four weeks. We examine the effectiveness of these different approaches to

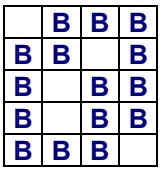
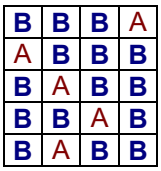
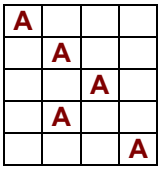
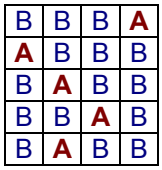
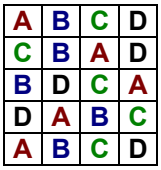
	Single Frequent *(no distracter)*	versus	***Infrequent Distracter***
			
Failure Rate	0.35%	χ^2 (1,540) = 3.3, p = 0.069	1.94%
Attempts Required	1.04 tries	**χ^2 (1,540) = 4.0, p = 0.044**	1.13 tries
Login Time	13.78 sec	**χ^2 (1,540) = 16.5, p < 0.001**	16.31 sec
Total Time	14.59 sec	**χ^2 (1,540) =17.1, p < 0.001**	19.20 sec

Figure 5: We examined *interference* for *frequent* password use by examining the impact of an *infrequent distracter* on a *single frequent* password. Participants in the *infrequent distracter* condition required more login attempts, more time per successful login, and more total authentication time.

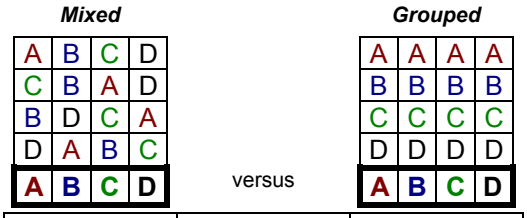
	Single Infrequent *(no distracter)*	***Frequent Distracter***	***Multiple Infrequent Distracters***
			
Failure Rate	1.45%	8.00%	15.23%
Attempts Required	1.17 tries	1.44 tries	1.68 tries
Login Time	18.29 sec	24.58 sec	29.79 sec
Total Time	22.32 sec	35.99 sec	47.35 sec

	Single Infrequent versus ***Frequent Distracter***	***Single Infrequent*** versus ***Multiple Infrequent Distracters***	***Frequent Distracter*** versus ***Multiple Infrequent Distracters***
Failure Rate	χ^2 (1,144) = 3.7, p = 0.054	**χ^2 (1,312) = 13.4, p < 0.001**	χ^2 (1,318) = 2.8, p = 0.093
Attempts Required	**χ^2 (1,144) = 4.6, p = 0.032**	**χ^2 (1,312) = 10.7, p = 0.001**	χ^2 (1,318) = 1.2, p = 0.266
Login Time	**χ^2 (1,137) = 7.9, p = 0.005**	**χ^2 (1,274) = 41.8, p < 0.001**	**χ^2 (1,275) = 8.8, p = 0.003**
Total Time	**χ^2 (1,137) = 8.4, p = 0.004**	**χ^2 (1,274) = 42.3, p < 0.001**	**χ^2 (1,275) =7.2, p = 0.008**

Figure 6: We examined *interference* for *infrequent* password use by examining the impact of a single *frequent distracter* and *multiple infrequent distracters* on a *single infrequent* password. Participants with *multiple infrequent distracters* were more likely to fail to authenticate. Both distracter conditions required more login attempts, more time per successful login, and more total authentication time.

training by analyzing data from the fifth week, when both conditions required accessing all four passwords. We hypothesized that the *mixed* condition would result in more failures, more required attempts, more time spent on a successful authentication, and more total time spent.

A chi-squared test of *failure rate* shows that the *mixed* condition resulted in significantly more failures ($\chi^2(1,142) = 4.7$, $p = 0.031$). The distribution of *attempts required* is non-normal, and a Mann-Whitney test finds that participants in the *mixed* condition required significantly more attempts ($\chi^2(1,142) = 5.7$, $p = 0.017$). After excluding failures from our temporal analyses, both *login time* and *total time* are non-normally distributed. Mann-Whitney tests show no significant difference for *login time* ($\chi^2(1,133) = 0.22$, $p = 0.636$) nor *total time* ($\chi^2(1,133) = 1.1$, $p = 0.291$). The lack of a difference in our time measures is difficult to interpret, but the more than fourfold difference in the failure rate is itself compelling evidence of the effect of the *grouped* training.



	Mixed	versus	Grouped
Failure Rate	11.29%	**$\chi^2(1,142) = 4.7$, $p = 0.031$**	2.5%
Attempts Required	1.56 tries	**$\chi^2(1,142) = 5.7$, $p = 0.017$**	1.19 tries
Login Time	24.27 sec	$\chi^2(1,133) = 0.22$, $p = 0.636$	26.88 sec
Total Time	34.50 sec	$\chi^2(1,133) = 1.1$, $p = 0.291$	33.14 sec

Figure 7: We examined the *training* of multiple passwords, comparing *mixed* training with *grouped* training. Participants in the *mixed* condition were more likely to fail to authenticate and required more login attempts.

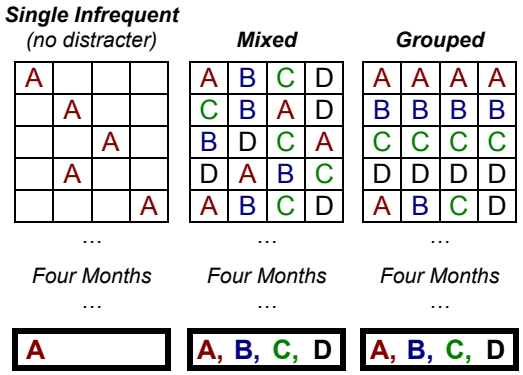
LONG-TERM RECALL METHOD AND RESULTS

Prior work has suggested that a single facial graphical password can be successfully recalled after significant periods of non-use [22]. We sought to examine the long-term recall of *multiple* facial graphical passwords, more specifically the impact of *interference* and *training*. Figure 8 illustrates our planned contrast of recall for participants in the *single infrequent* condition, the *mixed* condition, and the *grouped* condition. We hypothesized that the *single infrequent* condition would be easier than both *mixed* and *grouped*, with *mixed* also being more difficult than *grouped*, as indicated by all four of our measures.

We tested long-term recall by emailing participants four months after the end of our original five-week study. Prompts to authenticate were sent in separate email for each of a participant's passwords, and these emails were sent in a randomized order. Of the 69 participants in the three conditions we study here, 50 responded to our prompts.

Our results show a stark difference in the long-term recall of a *single infrequent* facial graphical password versus multiple infrequent facial graphical passwords with *mixed* training. Although every authentication participant in the *single infrequent* condition was successful, 14.3% of authentications in the *mixed* condition failed. A chi-squared test shows this difference is significant ($\chi^2(1,66) = 4.5$ $p = 0.035$). After excluding failures from this comparison, both *login time* and *total time* are non-normally distributed. Mann-Whitney tests show that participants in the *mixed* condition required significantly more *login time* ($\chi^2(1,59) = 5.2$, $p = 0.023$) and significantly more *total time* ($\chi^2(1,59) = 4.8$, $p = 0.028$) than those in the *single infrequent* condition.

Contrary to our expectation that the *grouped* condition would be more difficult than the *single infrequent* condition, our results show relatively little difference. Although participants in the *grouped* condition required significantly more *login time* ($\chi^2(1, 95) = 5.1$, $p = 0.024$)



	Single Infrequent (no distracter)	Mixed	Grouped
Failure Rate	0 %	14.29 %	0 %
Attempts Required	1.18 tries	1.63 tries	1.18 tries
Login Time	20.76 sec	31.71 sec	28.22 sec
Total Time	24.29 sec	47.27 sec	32.86 sec

	Single Infrequent versus ***Mixed***	***Single Infrequent*** versus ***Grouped***	***Mixed*** versus ***Grouped***
Failure Rate	**$\chi^2(1,66) = 4.5$, $p = 0.035$**	$\chi^2(1, 95) = 0$, $p = 1.0$	**$\chi^2(1,127) = 14.0$, $p < 0.001$**
Attempts Required	$\chi^2(1,66) = 2.2$, $p = 0.135$	$\chi^2(1, 95) = 0.04$, $p = 0.833$	**$\chi^2(1,127) = 5.3$, $p = 0.022$**
Login Time	**$\chi^2(1,59) = 5.2$, $p = 0.023$**	**$\chi^2(1, 95) = 5.1$, $p = 0.024$**	$\chi^2(1,120) = 0.3$, $p = 0.616$
Total Time	**$\chi^2(1,59) = 4.8$, $p = 0.028$**	$\chi^2(1, 95) = 3.5$, $p = 0.062$	$\chi^2(1, 120) = 0.98$, $p = 0.322$

Figure 8: We examined *long-term recall* by prompting participants to authenticate four months after they completed the main portion of our study. The impact of *interference* and the importance of *training* persisted.

and marginally more *total time* ($\chi^2(1, 95) = 3.5$, $p = 0.062$) than participants in the *single infrequent* condition, all authentications are successful in both conditions and there is no significant difference in the *attempts required* to authenticate ($\chi^2(1, 95) = 0.04$, $p = 0.833$). This is consistent with our result that the *grouped* condition had a significantly lower *failure rate* ($\chi^2(1,127) = 4.5$, $p < 0.001$) and significantly lower *attempts required* ($\chi^2(1,127) = 5.3$, $p = 0.022$) versus the *mixed* condition.

POST-STUDY QUESTIONNAIRE RESULTS

Responses to our post-study questionnaire show that many participants liked the idea of using facial graphical passwords: 41% said they would "definitely" use a facial graphical password system, 32% said they would "probably" use a facial graphical password system, and 27% said they would not use a facial graphical password system. Participant comments suggest a tension between the fact that participants generally were reasonably successful in using multiple facial graphical passwords versus the fact that some of their existing password management strategies (such as reusing passwords or writing passwords down) were inapplicable to this study. For example, one participant described a feeling that they were guessing:

> "*I found I often felt like I was guessing, but I usually guessed right, so I guess I remembered the right faces somewhere in the back of my mind. However, I didn't feel very comfortable with my choices*."

Another participant described a similar feeling that facial graphical passwords lack cues to help with remembering a password that has been partially forgotten:

> "*The one time I forgot my faces, I was totally [in a bad position]. It wasn't like I could remember one of my faces and use that as a clue to remember the rest. Each face seemed completely separate.*"

Some participants were surprised by their long-term recall of multiple facial graphical passwords. A participant from the *grouped* training condition commented:

> "*It's freakin' amazing that I remember all these!*"

None of the participants attempted to record their study passwords, but 29% reported they would use screen captures, sketches, or notes to attempt to document a facial graphical password outside of a study context.

DISCUSSION

We have presented the first study of *multiple* graphical passwords to systematically examine the effect of *frequency* of access to a graphical password, the effects of *interference* resulting from interleaving access to multiple graphical passwords, and the effect of patterns of access while *training* multiple graphical passwords. The effects discussed throughout our results sections have a number of important implications for graphical passwords.

Our findings regarding *interference* show that field studies of graphical password systems are likely to overestimate ease of access if they do not study the realistic use of multiple graphical passwords. In our largest interference contrast, we saw that participants accessing four different *infrequent* passwords each week had a failure rate more than ten times greater than participants accessing a single *infrequent* password. People typically have a need for many more than four passwords, so we might expect the effects of *interference* to be even more dramatic in a widespread deployment of graphical passwords. We also note that the impact of *interference* was not limited to our most extreme contrasts. Even our most mild examination of *interference*, Figure 5's addition a single *infrequent distracter* to a *frequent* password, resulted in a marginally greater *failure rate*, significantly more *attempts required*, and significantly more *login time* and *total time* to authenticate. In contrast to typical studies that examine only a single graphical password, our findings underscore a need for more realistic evaluations of the use of *multiple* graphical passwords. Given our young and technically-experienced participant population, our *frequency* and *interference* results might be considered a sort of lower bound: we have shown that ease of authentication is significantly impacted by *frequency* and *interference*, and the size of this effect might be larger with more heterogeneous populations or less frequent access.

Our findings regarding *interference*, *training*, and *long-term recall* motivate future field studies examining how people typically acquire and learn new passwords. The *interference* associated with multiple *infrequent* passwords was greatly reduced by the *grouped* approach to training, an effect easily seen in both the fifth week of our study and four months later. Although prior work has examined how many passwords people typically have, our results show it is also important to better understand the rate at which people acquire new passwords and the extent to which people are likely to have opportunities to practice a new password. Informed by our results and future field studies, developers of graphical password systems will be able to study ease of access under more realistic *training* conditions.

In addition to informing more realistic evaluations of graphical password systems, our results have a number of implications for the adoption of graphical password systems. The effectiveness of our *grouped* training, for example, suggests that applications employing graphical passwords might consider encouraging a burst of initial usage over the course of the week following creation of a new password. As another example, our study design ensured there was no overlap of faces appearing in any of a participant's multiple passwords nor their distracter faces, but our *interference* results and participant comments in our post-study questionnaire suggest that it would be quite problematic if the same face appeared in multiple contexts. More generally, both developers and adopters of graphical password systems should be wary of default image databases shipped with graphical password systems and the risk of increased *interference* if those databases are used on multiple sites. Finally, our results show that the *time* required to authenticate can be significantly impacted by *frequency*, *interference*, and *training* even when the *failure*

rate is not. Designers considering the role of graphical passwords in applications therefore need to be sure they have realistic estimates of the time that will be required to authenticate using a particular graphical password system. Estimates that ignore the effects of *frequency*, *interference*, and *training* may be unrealistically optimistic and may lead to unacceptably cumbersome designs under realistic use.

Finally, our demonstration of the effects of *frequency*, *interference*, and *training* on multiple graphical passwords provides a new basis for comparing proposed graphical password systems. We chose to study facial graphical passwords because of their commercial deployment [17] and their use in prior research [4, 9, 21], but similar studies should be conducted for other graphical password systems. It is possible, for example, that other cognometric, locimetric, or drawmetric schemes are less susceptible to *interference*. If a graphical password system were found that performed extremely well in our *multiple infrequent distracters* condition, for example, this would be a strong indication that the system retains its ease of access even with *interference* from multiple graphical passwords.

CONCLUSION

We have presented the first study of *multiple* graphical passwords to systematically examine the effects of *frequency, interference*, and *training*. In contrast to prior work's examination of a single graphical password, our results underscore the need for more realistic evaluations of the use of *multiple* graphical passwords, have a number of implications for the adoption of graphical password systems, and provide a new basis for comparing proposed graphical password systems.

ACKNOWLEDGEMENTS

We would like to thank our study participants. This work was supported in part by National Science Foundation grant IIS-0812590 and an Alfred P. Sloan Research Fellowship.

REFERENCES

1. Adams, A. and Sasse, M.A. Users are not the enemy. *Communications of the ACM*, (CACM Dec 1999), 40-46.
2. Adams, A., Sasse, M.A., and Lunt, P. Making passwords secure and usable. *Proceedings of HCI on People and Computers XII*, (HCI 1997), 1-19.
3. BBC News. *UN warns on password 'explosion'*. http://news.bbc.co.uk/2/hi/technology/6199372.stm.
4. Brostoff, S. and Sasse, M.A. Are Passfaces™ more usable than passwords? A field trial investigation. *Proceedings of HCI on People and Computers XIV*, (HCI 2000), 405-424.
5. Chiasson, S., Biddle, R., and van Oorschot, P.C. A second look at the usability of click-based graphical passwords. *Proceedings of the Symposium on Usable Privacy and Security*, (SOUPS 2007), 1-12.
6. Davis, D., Monrose, F., and Reiter, M. On user choice in graphical password schemes. *Proceedings of the Conference on USENIX Security Symposium*, (2005), 11-11.
7. DeAngeli, A., Coventry, L., Johnson, G., and Renaud, K. Is a picture really worth a thousand words? Exploring the feasibility of graphical authentication systems. *International Journal of Human-Computer Studies*, v. 63, n. 1-2 (2005), 128-152.
8. Dhamija, R. and Perrig, A. Déjà vu: A user study using Images for Authentication. *Proceedings of the Conference on USENIX Security Symposium*, (2000), 4-4.
9. Dunphy, P., Nicholson, J., Olivier, P. Securing Passfaces for Description. *Proceedings of the Symposium on Usable Privacy and Security*, (SOUPS 2007), 24-35.
10. Ensor, B. How Consumers Remember Passwords. *Forrester Research Report*, June 2, 2004.
11. The Face of Tomorrow Face Dataset. http://www.flickr.com/photos/istanbulmike/sets/72157594201837268/.
12. Florencio, D. and Herley, C. A large-scale study of web password habits. *Proceedings of the International Conference on World Wide Web*, (WWW 2007), 657-666.
13. Gaw, S. and Felten, E. Password management strategies for online accounts. *Proceedings of the Symposium on Usable Privacy and Security*, (SOUPS 2006), 44-55.
14. Ives, B., Walsh K.R., and Schneider, H. The domino effect of password reuse. In *Communications of the ACM*, (CACM Apr 2004), 75-78.
15. Moncur, W. and Leplâtre, G. Pictures at the ATM: exploring the usability of multiple graphical passwords. *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*, (CHI 2007), 887-894.
16. Morris, R. and Thompson, K. Password security: A case history. *Communications of the ACM* (CACM Nov 1979), 594-497.
17. Passfaces™. http://www.realuser.com/
18. Rock, I., & Engelstein, P. (1959). A study of memory for visual form. *American Journal of Psychology* (1959), 72, 221–229.
19. Standing, L. Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology* 25 (1973), 207–222.
20. Tari, F., Ozok A.A., and Holden, S.H. A comparison of perceived and real shoulder-surfing risks between alphanumeric and graphical passwords. *Proceedings of the Symposium on Usable Privacy and Security*, (SOUPS 2006), 56-66.
21. Valentine, T. An evaluation of the Passfaces™ personal authentication system. *Goldsmiths College Technical Report,* 1998.
22. Valentine, T. Memory for Passfaces™ after a long delay. *Goldsmiths College Technical Report*, 1999.
23. Wiedenbeck, S., Waters, J., Birget, J.C., Brodskiy, A., and Memon, N. PassPoints: Design and longitudinal evaluation of a graphical password system. *International Journal of Human-Computer Studies*, v. 63, n. 1-2, (2005), 102-127.