Free-Form Gesture Authentication in the Wild

Yulong Yang[†], Gradeigh D. Clark[†], Janne Lindqvist[†], Antti Oulasvirta[*]
[†]Rutgers University, [*]Aalto University
{yulong.yang, gradeigh.clark, janne.lindqvist}@rutgers.edu, antti.oulasvirta@aalto.fi

ABSTRACT
Free-form gesture passwords have been introduced as an alternative mobile authentication method. Text passwords are not very suitable for mobile interaction, and methods such as PINs and grid patterns sacrifice security over usability. However, little is known about how free-form gestures perform in the wild. We present the first field study (N=91) of mobile authentication using free-form gestures, with text passwords as a baseline. Our study leveraged Experience Sampling Methodology to increase ecological validity while maintaining control of the experiment. We found that, with gesture passwords, participants generated new passwords and authenticated faster with comparable memorability while being more willing to retry. Our analysis of the gesture password dataset indicated biases in user-chosen distribution tending towards common shapes. Our findings provide useful insights towards understanding mobile device authentication and gesture-based authentication.

Author Keywords
Mobile; authentication; free-form gesture; field study; ESM.

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User Interfaces—*Input devices and strategies*; K.6.5 Security and Protection: Authentication

INTRODUCTION
Smartphones today store a large amount of users' personal and sensitive data (e.g. financial information) [52]. Allowing only authorized access to this data is of increasing importance, and highlights the need for research on usable and secure smartphone authentication. Text passwords, although still the most used authentication method, have been criticized by many [5, 6, 1], and found to be particularly unsuitable for mobile use [43, 64, 60]. Other existing methods (PIN, grid-based pattern, biometrics) on mobile devices suffer from various shortcomings: limited password space [58], susceptibility to shoulder surfing [18, 61], easily crackable [7, 8, 55], slow entry [17] and harmful for privacy [15].

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858270

Free-form gesture passwords have been recently proposed as an alternative for mobile user authentication [51, 11]. Free-form gesture passwords allow users to draw any shape or pattern on a blank touchscreen display using one or more fingers. Several previous studies demonstrated that, for both mobile and non-mobile use, free-form gesture passwords can be secure and memorable [51, 57]. They potentially provide a large password space because of their free-form nature. Mobile interaction is found to be fragmented, frequent and short-term [44, 28]. Since gesture-based interaction conforms to the form factor of mobile devices [47] and is faster than typing, it is more suitable for authentication than text. In addition, when used as a password, free-form gestures improve memorability with the help of visual learning effects [45] and motor memory [24].

However, most work on gesture passwords so far has been carried out in laboratories [51, 48, 16, 26], leaving their performance in the wild as an open research question. Field studies are important for understanding the user-chosen distribution of gesture passwords in realistic settings and how usable and memorable those could be.

In addition, previous work has focused on using gesture-based authentication for a single account or phone unlocking [51, 50, 26, 48, 16], and has not considered it for multi-account configurations. However, people manage multiple accounts at the same time in reality [25, 31]. Previous work also shows how multi-account settings affect the authentication process. For example, a study showed that multi-account interference significantly impacts the ease of authentication of facial graphical passwords [21]. Therefore, it is crucial to explore how gesture passwords would be different under the multi-account context.

In this paper, we report the results of the first field study using free-form gestures as a mobile authentication method. We used text password as a baseline comparison given that it is still the most commonly deployed authentication method [6]. We leveraged the Experience Sampling Method (ESM) [13] to implement the field study. ESM uses a signaling device to prompt participants to respond to tasks or questionnaires [12, 32]. Normal field studies usually hand over control of the experiment to participants entirely. As a result, it is hard to assess aspects such as memorability of a password because the frequency and engagement of each participant differs. With ESM, researchers are able to control when and where to notify participants to complete tasks, thus ensuring comparable contributions among participants. Meanwhile, it provides better ecological validity over laboratory studies because it allows participants to perform tasks during their regular rou-

tines and in real-world settings [10, 13]. There have already been ESM-type studies on mobile interactions (e.g. [10, 46]) and device unlocking [28], but none on gesture password use.

In our study, participants were randomly assigned to either text or gesture group. We then installed our application on their devices and throughout the study they received pre-scheduled native notifications on their smartphones. Each notification represented either a password generation or password recall task. Each task was designed with various intervals and each participant completed such tasks for eight virtual accounts totally.

Overall, our 91 participants generated 347 text passwords and 345 gesture passwords with 2002 completed log-in tasks. Our findings contextualize the existing research on gesture passwords as well as challenge previous findings from lab studies. We found that gesture passwords demonstrated better usability over text passwords. In general, participants with gesture passwords spent less time both generating new passwords and logging in. The difference between the two groups was statistically significant under multi-account interference. In addition, participants with gesture passwords were more willing to retry before giving up. Text and gesture passwords showed similar memorability, but gesture passwords performed better under multi-account settings in the short term. We also proposed a metric to compare the security of the two passwords types uniformly. We found that the collected gesture passwords carried comparable and possibly higher entropy than the text passwords.

We also present the analysis of the first gesture password dataset from the field. We found participants preferred shapes (49.28%) and letters (24.07%) as their passwords, with 15.36% of the gesture passwords being symmetric. Participants also preferred single-finger gestures (93.62%) over multi-finger ones.

In summary, this paper contributes by showing that free-form gestures improve over text passwords in mobile contexts. More precisely, we make the following major contributions:

1. This is the first field study on *memorability* and *usability* of free-form gesture passwords.
2. Our results indicate that free-form gesture passwords are more resilient to multi-account interference than text passwords.
3. Our results show that free-form gesture passwords provide better mobile usability than text passwords.
4. We discovered that participants tend to choose common shapes, letters, and account-specific gestures as passwords.

RELATED WORK

Free-form gesture passwords have been proposed as mobile authentication method and evaluated to be secure and memorable in the lab. Sherman et al. developed an information-theoretic metric and conducted a lab study showing that single-finger gestures and gestures with many hard angles and turns demonstrated better security [51]. Another lab study, focusing on user interfaces and not authentication, found that user-defined gestures demonstrated better memorability than pre-defined ones [41]. Moreover, studies indicated free-form gestures were resistant against a major threat on mobile platforms: shoulder surfing [51]. There are also studies that combine gestures with other factors for authentication: simple strokes on the mobile device itself [16], continuous gesturing [26] and tapping actions [65]. In addition to mobile device gestures, researchers have explored mid-air gestures for authentication [57].

There have been many field studies on other authentication schemes. A field study on recognition-based graphical passwords found that, among other results, ease of authentication was significantly impacted by multi-account interference [21]. Alt et al. found that 51% of image-based passwords from the field could be predicted by human attackers [2]. Egelman et al. showed that the effect of strength meters on passwords differed for different contexts with a field study followed after a lab study [20]. A field study comparing the usability of the Android grid-based pattern unlock to PINs has indicated that participants preferred to use the former despite the latter having higher input speed and fewer errors [62]. Another study focused on designing and implementing strength meters for pattern unlock and found it improve the security [53]. A week-long field study on different graphical password schemes (free-recall, cued-recall, and recognition) concluded that both of the last two were superior to free-recall, though users preferred the recognition scheme despite longer login times [56].

There is, however, limited literature applying ESM to mobile authentication studies. One study utilized ESM to capture participants' perceptions towards unlocking behaviors, revealing reasonings behind leaving a phone unlocked [28]. Another similar self-reporting methodology, diary studies, has been used in recent research. One diary study on the cost of password policies had 32 staff members record 196 password events over one week [34]. Another study asked participants to record password events when they log into their accounts using desktop computers or laptops [31]. A diary study showed that authentication tasks lowered the productivity of employees in an organization [49].

In summary, it is clear that two major gaps exist in the research of free-form gesture passwords: (1): Little is known about how they perform in the field; (2): No previous work considered the scheme under multi-account contexts.

METHOD

Our experiment focused on comparing the use of two types of passwords in the field. We describe our method below.

Participants

We recruited participants through fliers and mailing lists. Participants were required to be at least 18 years old, have familiarity with and own an Android smartphone. All participants received a $30 gift card as compensation, and enrolled in a raffle for three $75 gift cards. The study was approved by the Institutional Review Board at Rutgers University.

We recruited 110 participants. Three participants withdrew during the study, and another 16 participants were excluded from our analysis as they did not complete at least half of the

tasks that the study required. We excluded them to reduce bias in our analysis. This reduced our sample to 91 participants. Our participation rate was above average as compared to similar studies [40, 62, 21]. The remainder of this paper focuses on these 91 participants.

Our participants' ages were from 18 to 52 (mean = 23.03, SD = 7.01, Mdn=21). 47 participants were male and 44 were female. 56.04% of them were college students, 23.08% were graduate students, 6.60% were engineering or IT professionals, and 4.40% worked in management or finance.

Our participants reported to be experienced and frequent smartphone users: 82 (90.11%) have used smartphones for more than one year and 59 (64.84%) for more than three years; 83 (91.21%) spend at least two hours per day on smartphones and 42 (46.15%) spend four hours or more per day.

We also collected form factor data — screen resolution by number of pixels. The result indicates most participants used modern and up-to-date devices: 65.96% of them used a smartphone of resolution 1080×1920, 17.02% used 720×1280, and 17.02% used others.

Experiment Design

Our study consisted of two main tasks: creating passwords for a specific virtual account and logging into the account with the created passwords. In the experiment, we focused on comparing two types of passwords and we varied the number of accounts and recall interval as well.

Password type is a between-group variable aimed at comparing text and gesture passwords. Each participant was randomly assigned into either the text or gesture group. We performed randomized assignments in a way such that each group had equal or similar sample sizes. The main reason to choose between-group over within-subject design was to avoid interference effects between the two types. We adopted cues from previous studies on multiple password interference [9, 42, 21]. We chose text password as a baseline comparison, because one of our objectives was to study real-world multi-account interference with gestures; scenarios such as device unlock (pattern unlock, PIN) do not require managing multiple accounts.

The number of accounts refers to how many accounts participants had to manage during the study. We designed this for two purposes: (i) to study the multi-account interference of the passwords, and (ii) to achieve better ecological validity since people usually have multiple accounts in the real world [25]. Each participant was asked to create and recall passwords for two different account sets. All accounts were created for the purpose of this study. The first set contained two virtual accounts: online banking and social network. The second set had six accounts: email, online gaming, online dating, shopping, online course, and music streaming. We chose common services that were easy to understand and distinguish between, as opposed to something more generic (e.g. "account A"). Accounts were differentiated from each other by their names, logos, and colors.

The log-in time interval is the time between the log-in and the creation task: immediate log-in tasks occurred one hour after the completion of a creation task, short-term and long-term tasks occurred one day and one week later. This design was intended to study the effect of time on metrics such as memorability.

Our tasks followed the process where people log in to their online services remotely. In practice, to securely store gesture passwords for the process, we could utilize existing work that solved similar issues such as fuzzy vaults [35].

Experience Sampling Method

We incorporated the Experience Sampling Method (ESM) in our field study. ESM is a research methodology where participants perform tasks at different points of time during their daily lives. It is usually used to capture the subjective experience of participants in a natural environment [12, 3, 28].

We leveraged ESM in our design in two ways. First, following the idea behind ESM wherein participants are alerted multiple times per day for self-reporting [12], we scheduled tasks to arrive at different times of the day. Creating or logging in with multiple passwords at the same time is far from the actual use case concerning passwords. Asking participants to use unique passwords at different points of time improved the ecological validity of the study. This is also suitable for simulating mobile authentication, where log-ins are frequent and can occur at any time [28, 19, 53, 39, 25]. We discuss the schedule of our study in the Procedure section below.

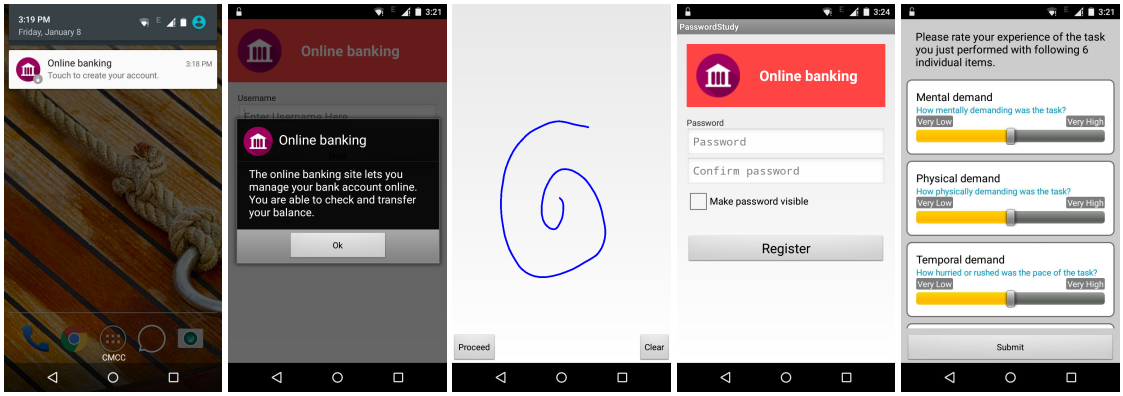
ESM also emphasizes that participants react at the moment they are alerted in order to collect precise data [3]. We incorporated this concept by setting our tasks to expire one hour after they arrived. We gained better control over the field study as participants had to react within the scheduled time frame. The expiration window also helped maintain the log-in time interval. Different types of log-in tasks (immediate, short-term, long-term) were differentiated from each other by the time of their arrivals. Without the expiration window, participants might delay responding to tasks. This means that the original schedule would be disrupted and the preset log-in time intervals would become meaningless.

Apparatus

We built an Android application to install on our participants' devices. It was responsible for (i) notifying participants based on a preset schedule, and (ii) allowing participants to complete tasks. The application had two versions, differing only by the type of password it supported. Participants installed only one version based on which group they were in. Figure 1 shows sample screen-shots of the application.

The notification generated by our application was native and directly viewable on the device. It remained alive until either participant completed the corresponding task or it expired. It was generated offline on the phone, so Internet or cellular access was not required. Figure 1a shows a sample screenshot of our task notification.

Tapping the notification brought participants to the user interface for either a password generation or log-in task.



(a) Notification example **(b) Account description interface** **(c) Gesture password input interface** **(d) Text password input interface** **(e) TLX form**

Figure 1: Sample screen-shot of the application. From left to right: notification, account description dialog, gesture password input interface, text password input interface and TLX form. One application had only one password input interface – either gesture or text password. Accounts were differentiated from each other by their names, logos, colors and descriptions. Notification appears in the notification drawer to alert participants about incoming tasks.

Both tasks were a common two-step authentication process wherein a participant first input their username followed by their password. Every task was followed by a validated NASA subjective Task Load Assessment (TLX) form [29] to fill out. The form is designed to estimate workloads subjectively using six individual factors [29]. Participants were asked to give each factor a score based on their experience of the assessed task, with each score ranging from 0 to 100.

In prior studies, participants have been alerted by email [21, 62, 56, 14]. The email usually contains a link to a website where participants then perform tasks. However, our design brings one major advantage that better fits mobile authentication usage: it is not limited by participant contexts. Our participants could perform the task on their smartphones without any need for desktops, internet, or cellular access. As mobile authentication can occur in various contexts, our design adequately simulated the process.

Gesture Password Authenticator

Our gesture password authenticator is a modification of Protractor [38] that was extended [51] for multi-touch gesture support. Protractor is a gesture recognizer that measures the cosine distance between a stored template of a gesture password and an input, and uses the reciprocal of that as a *similarity score*. Discussion of the design and implementation is beyond the scope of this paper, and has been explained in exhaustive detail in previous work [38, 51].

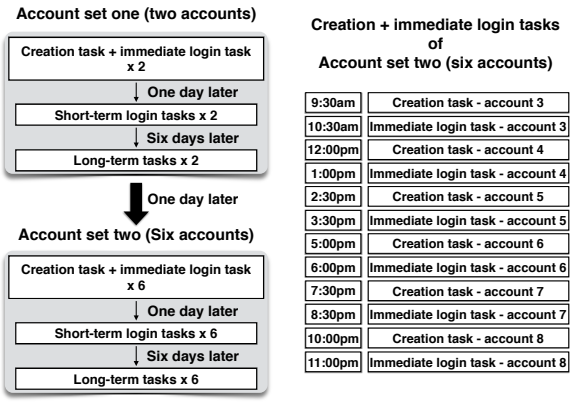
Authentication success is determined by whether the computed score is greater than a threshold. Selection of the threshold has not been examined in great detail nor are there any defined heuristics for computing it in the literature. As such, we derived it based on existing gesture data. In a previous study, we collected a dataset of over 60 distinct free-form gestures and 1200 generated trials [51]. The average score of all trials was 2 after rounding, therefore we chose 2 as our threshold. This empirical selection represented an average case of over 1200 separate scores and was reasonably low enough to authenticate participants without requiring a taxing amount of accuracy from them, while high enough to prevent obviously incorrect passwords from being accepted.

Procedure

At first, participants were introduced to the study and asked for consent to participate in the experiment. After consenting, we installed the application on their smartphones and demonstrated how it works with a testing device. We also informed participants that passwords they generate should be secure, easy to memorize, and difficult for others to guess. Moreover, we emphasized that they should not use their real passwords, nor use any password managers to help memorization (including writing the password down).

For the next two weeks, participants performed tasks on their devices in their daily lives. This process is illustrated in Figure 2a. In the first week, participants performed tasks for the first account set (of two accounts). On the first day of the week, participants were asked to create usernames and passwords for the two accounts at different times. Each creation task was followed by a corresponding log-in task one hour later. For each account, short-term and long-term log-in tasks arrived one day and one week after the creation task, respectively. The actual order of tasks of those accounts was different for each participant. It was based on a latin square arrangement in order to avoid potential bias from scheduling. In the second week, participants went through a similar process for the second account set (of six accounts).

To properly distribute tasks within one day, our considerations were two-fold: (1) the schedule should not disturb participants' daily life too much; (2) it should cover the majority of time during which people are likely to use passwords. Previous work indicates most password usage occurs from 6:30 AM to 10:00 PM [25]. Our range of time to schedule tasks was 9:30 AM to 11:00 PM. We shifted and stretched the range to fit the normal schedule of most people. Inside the time range, we tried to distribute tasks evenly into equal-length time blocks. Figure 2b shows a typical schedule of a day in our study.



(a) The 2-week process of our study. **(b) A sample schedule of creation + immediate login session for account set two.**

Figure 2: The left figure shows the process of the entire study. The right figure shows a typical schedule of for a day of creation + immediate login tasks for one participant.

Participants were allowed to withdraw under any circumstances without penalty. After two weeks, we invited them back for a brief interview and to receive their compensation.

Password Analysis

We performed password cracking attacks to analyze text password security. We used the popular GPU-based password cracker oclHashcat 1.36 [30] to generate rule-based attacks. The cracker generates guesses by applying rules to modify words in the dictionary; for example, one rule could be to capitalize the first character in every word.

We generated three attacks. The first two attacks used rule sets that come with the software by default, called "basic64" and "generated2". The third one used the rule set designed by KoreLogic [33]. It is a subset of rules they used to generate passwords for DEF CON's "Crack Me If You Can" password-cracking contest in 2010 [36]. It has been found to be effective for password cracking [54]. All attacks used the same input dictionary, a shuffled combination of different wordlists that included Google 1-gram English dataset [27], UNIX dictionary [37], RockYou leaked password dataset, and phpbb leaked password dataset. It contained 38M unique words. We followed the state-of-the-art cracking techniques from recent literature [59] so that the results are comparable.

We also analyzed unique passwords we collected. While unique text passwords could be easily determined by comparing the text of two passwords directly, for gestures, we relied on the score computed by our authenticator. We started with the unique gesture set as empty and iterated over all our gesture passwords. For each gesture, the authenticator computed a score of it and every other gesture in the set. If the max score of them was smaller than our threshold, we determined it as a unique password and added it to the unique gesture set.

Group	Size	Creation completed	Log-ins completed
Total	91	692 (95.05%)	2002 (91.67%)
Text	44	347 (98.58%)	960 (90.90%)
Gesture	47	345 (91.76%)	1042 (92.38%)

Table 1: Study statistics overview. "Creation completed" lists the number of passwords generated by each group. "Log-ins completed" shows the number of log-in tasks completed by each group. The percentage after each number indicates the completion rate of the particular item. The completion rate of an item is the percentage of designed tasks that were eventually completed by participants.

Security comparison

To compare the security of two types of password, we calculated the random entropy for them. The equation for calculating random entropy for text passwords is $H = L \times \log(N)$, where L is the password length, and N is the amount of possible characters. To model gestures similarly, we treated a free-form gesture as points on the touchscreen connected through one stroke, and the screen as many equal-size cells. A point belongs to one cell given its location. We define L as the number of points, and A and B to be the number of cells horizontally and vertically we segment the screen into, respectively. Then, the number of cells is the equivalent of the possible character size in text password case (N=AxB). In this model, the more points a gesture has and the more cells a touchscreen is split into, the more fine-grained a gesture could be, and therefore higher entropy it contains.

Statistical Tests

For our categorical data, such as the log-in success rate, we used chi-square tests. We used the non-parametric equivalent of t-test, Wilcoxon rank sum test [63], to compare the continuous data of two password groups. We chose $p < .05$ to indicate whether the test result is statistically significant. When multiple comparisons existed, we used the adjusted p value based on Bonferroni correction. Bonferroni correction is a common method to control familywise error rate when dealing with multiple comparisons [23]. For easier reading, we used tables to report the tests for multiple comparisons.

RESULTS

Table 1 shows an overview of the amount of generated passwords and attempted log-ins. 91 participants generated 692 passwords and performed 2002 password log-in tasks with a completion rate of 95.05% and 91.67%, respectively.

The average response time to a single task was 7.71 minutes (Mdn=1.27). Response time is the duration of time from when a task is signaled to the time when the participant respond to it. 75% of our participants reacted to a task within 10.45 minutes.

Below we first present the result of our creation tasks, and then log-in tasks.

Creation Tasks

Duration

We found that the gesture group took less time to generate a new password than the text group when the number of accounts equaled six, as Figure 3 shows. The creation duration was calculated as the average time needed to create a password for one account. In two-account settings,

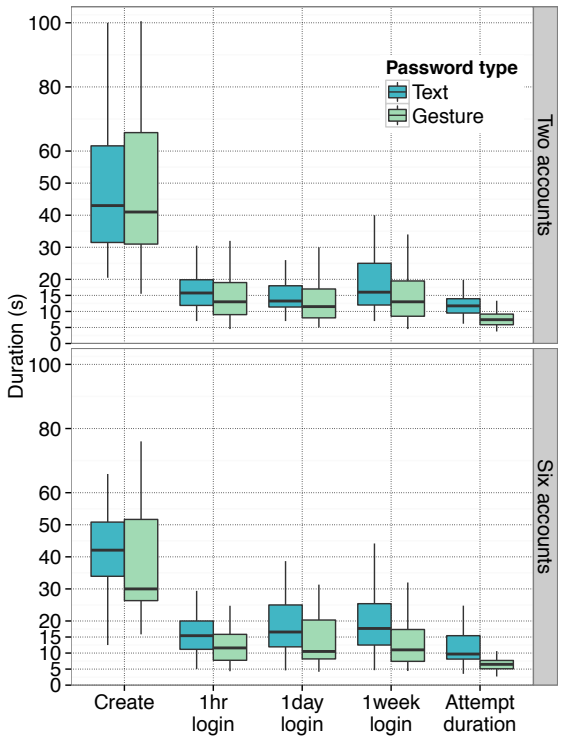


Figure 3: Duration data, including creation duration, log-in duration and attempt duration. Outliers are removed for clear visualization. The log-in duration is the average time needed to log in to one account successfully. The attempt duration is the time needed to perform one log-in attempt. The figure is broken down by account settings. One-hour (1hr) log-in, one-day (1day) log-in and one-week (1week) log-in corresponds to immediate, short-term and long-term log-in tasks, respectively.

the text group spent an average 58.56 seconds (Mdn=43) to create one password, while the gesture group spent 69.43 seconds (Mdn=41); when the number of accounts increased to six, the same task took the text group 76.38 seconds (Mdn=42.08), but only 44.04 seconds (Mdn=30) for the gesture group. According to a Wilcoxon rank sum test result, the text group used significantly more time than the gesture group in the six-account setting ($W = 1281.5, p = .0498, r = -.2056, 95\%\ C.I. = [3.41 \times 10^{-5}, 14.5]$). The two groups took similar times to generate a password in the two-account setting ($W = 1088, p = .6709, r = -.04$); no statistical significance was seen. The confidence interval indicates that the text group could spend as much as 14.5 seconds more than the gesture group to create one password.

Text Passwords Created

Our study generated 347 text passwords of which 209 were unique. We tested the strength of the passwords with the three cracking attacks described earlier in the Method section. Table 2 shows the results of general statistics and the result of the attacks. Two of them cracked more than half of the passwords. The result is similar to that of the weakest category of passwords being cracked in a recent study [59].

	Mean (SD)
Length	9.52 (3.07)
Lowercase	7.35 (3.21)
Uppercase	0.36 (0.78)
Digit	1.64 (1.79)
Symbol	0.16 (0.38)

Attack	# Guesses	Cracked (%)
Best64	3×10^9	40.19
Generated2	2.5×10^{12}	61.72
KoreLogic	1.6×10^{15}	68.42

Table 2: Text passwords' number of characters (left), and result of cracking attacks (right).

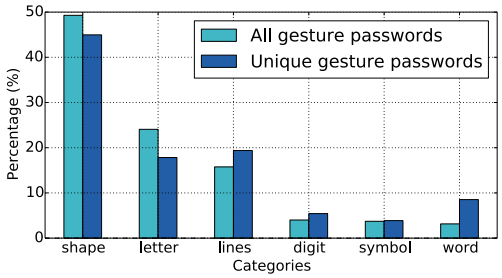


Figure 4: The categories of all gesture passwords created by the participants. The analysis shows the count of both all gestures passwords and only unique gesture passwords.

Free-form Gesture Passwords Created

Our study collected 345 gesture passwords overall, and 150 of them were unique. Among them, 22 were drawn with multiple fingers, and 53 were symmetric.

To better understand the collected passwords, we grouped them into six categories based on our observations. "Shapes" consisted of all passwords that were about real or virtual objects and geometric shapes. "Letters" contained gestures with letters and initials. "Symbols" had gestures that used common symbols that appear in e.g. computing or math. "Digit" gestures were those made of single digit numbers. "Lines" contained gestures that were either single line or a simple combination of several lines – these gestures were mostly abstract and did not refer to any obvious objects. Gestures in the "Words" group contained either words or signatures.

We found most passwords were "Shapes" and "Letters", as shown in Figure 4. We categorized our gesture dataset by both (i) all passwords, and (ii) unique passwords. Overall, the dataset containing the most passwords was the "Shapes" category (49.28%). The second and third most popular categories were "Letters" (24.07%) and "Lines" (15.76%).

Figure 5 shows the five most popular gesture passwords in the category "Shapes." Most of our popular gesture passwords were common shapes or objects, such as stars and squares. One of them, what we called "one-stroke" (see Figure 5e), could be due to our gesture authentication system. Our system required participants to draw the gesture in one stroke without lifting fingers; according to some participants, they chose "one-stroke" because it was easy to draw in one stroke.

Security comparison

With the metric we proposed in the Method section, we calculated the entropy a gesture password could contain given different N (number of cells on the screen), since L (number of points per gesture) is fixed to 16 in our case. The result is shown in Figure 6. The figure also includes a boxplot of entropy of our text passwords for comparison.

(a) First (b) Second (c) Third (d) Fourth (e) Fifth

Figure 5: The five most popular gesture passwords in "Shapes" category. From left to right, they are: star, square, heart, triangle, and one-stroke, respectively. They take up 37.21% of the entire password dataset.

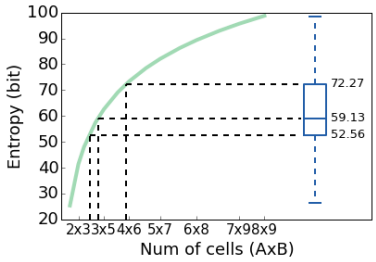


Figure 6: The curve displayed is the entropy of gesture passwords varied by the number of cells. Number of cells of a screen is defined as the number of cells horizontally (A) times number of cells vertically (B). The boxplot represents the entropy of our text passwords. The edge of the box represents first and third quantiles respectively, and the bar inside the box represents the second quantile (median). Values to the right of the box shows the entropy of each quantile.

According to the graph, the median of our text password entropy is 59.13, and is close to that of gesture passwords when we consider the touchscreen as a 3x5 grid of cells. It also shows that first quantile (25%) of text password entropy maps between 2x3 and 3x5 grid for gestures, and the third quantile (75%) maps to a grid of 4x6 cells.

Log-in tasks

Log-in Success Rate

We found the log-in success rate of the two groups to be similar. The success rate is the number of successful tasks divided by the total number of tasks across all participants. The overall success rates were 88.53% and 89.60% for the text group and the gesture group, respectively. Table 3 shows the success rate of logging in after one hour, one day and one week.

The table shows that the gesture group performed slightly better than the text group in most of the log-in tasks. However, we applied chi-square tests and found no statistically significant difference between the rate of two groups in any pairs shown in the table.

Duration

Successful log-in duration is the time participants needed to log in to a certain account successfully. We found that the gesture group spent less time in general than the text group in order to log in successfully. Overall, it took the text group 18.62 seconds (Mdn=15.8) on average and 16.49 seconds (Mdn=11.5) for the gesture group to log in to one account. Figure 3 shows the duration of each log-in session.

A Wilcoxon rank sum test showed that when number of accounts was six, the effect of password type on successful log-in duration was statistically significant for all three log-in sessions (see Table 4 for results). In particular, the time used by the text group increased as the number of accounts increased, while that of gesture group was relatively constant.

	Two accounts		Six accounts	
Log in after	Text	Gesture	Text	Gesture
One hour	97.56%	98.90%	97.59%	97.05%
One day	91.77%	93.55%	83.20%	83.57%
One week	88.24%	87.50%	80.42%	84.59%

Table 3: Success rate of two password types of each log-in task. Log in after one hour, one day and one week corresponds to immediate, short-term and long-term log-in tasks, respectively.

In addition, as time elapsed, the effect of password type on log-in duration was stronger. This is illustrated by the increase of the confidence interval of the difference in duration. The interval increased as the log-in task was further away from the creation task (see Table 4). After a week, participants with text passwords could spend two to eight seconds more than the gesture group in order to log into one account.

Attempts

We then looked at the number of attempts tried in each log-in task. Participants were allowed to retry unlimited times for any log-in task as long as it did not expire.

Overall, two groups retried similar times before they could successfully log in. On average, it took the text group 1.66 (Mdn=1) attempts, and the gesture group 2.44 (Mdn=1.5) attempts to successfully log in to one account. Table 5 shows detailed results. The attempts tried by the gesture group were slightly higher than that of the text group. However, the result of the Wilcoxon rank sum test shows that the effect of the password type on the number of attempts was statistically significant only when participants tried to log in after one hour in the two-account setting.

On the other hand, the gesture group was found to be faster when performing a single attempt. On average, it took the text group 11.33 seconds (Mdn=10.5) to perform a single log-in attempt, as compared to 7.71 seconds (Mdn=6.53) for the gesture group. We calculated it by dividing the duration of each log-in task by the number of attempts performed in that task. Because we only study the password usage, we removed the duration of inputting username from this calculation. Figure 3 shows the attempt duration of each log-in task.

To compare the attempt duration, we did a Wilcoxon rank sum test, and Table 6 shows the result of each pair-wise comparison. In all log-in tasks, the gesture group performed much faster than the text group in a single attempt. The confidence intervals show that the gesture group could login two to five seconds faster than the text group in a single attempt.

Errors

In our study, every time participants made a failed attempt, they generated an error. We also allowed them to give up on a log-in task at any time. In this subsection we look at the errors made by them and log-ins they eventually gave up on.

Overall, the gesture group generated more errors: 47 of them generated 1560 errors, while 44 participants in the text group failed 816 times. Half of the errors occurred in log-in tasks after one week (text: 52.21%, gesture: 46.92%).

We then categorized the type of errors made by both groups, which is shown in Figure 7. "Wrong account" referred to the

Account set	Log in after	Median (s) Text	Median (s) Gesture	Mean (s) Text	Mean (s) Gesture	Wilcoxon rank sum test W	p	r	95% C.I.
Two accounts	One hour	15.75	13.00	17.15	16.48	1267	.0647	-.1937	$[-3.29 \times 10^{-5}, 5.00]$
	One day	13.25	11.50	15.67	19.26	1234	.1120	-.1666	$[-0.50, 4.50]$
	One week	16.00	13.00	18.77	16.22	1220	.0350	-.2210	$[0.50, 6.50]$
Six accounts	One hour	15.41	11.60	17.20	14.58	1382	.0058*	-.2889	$[1.13, 6.07]$
	One day	16.57	10.50	21.84	15.45	1361	.0049*	-.2949	$[1.55, 7.70]$
	One week	17.67	11.00	21.17	16.92	1396	.0019*	-.3256	$[2.00, 8.33]$

Table 4: Successful log-in duration (seconds) of two password groups. Log in after one hour, one day and one week corresponds to immediate, short-term and long-term log-in tasks, respectively. The Bonferroni-corrected threshold p-value is .0083. The result shows the gesture group spent much less time to log in than the text group when the number of accounts was six.

Account set	Log in after	Median Text	Median Gesture	Mean Text	Mean Gesture	Wilcoxon rank sum test W	p	r	95% C.I.
Two accounts	One hour	1.00	1.00	1.24	1.91	755.5	.0081*	-0.28	$[-0.50, -1.60 \times 10^{-5}]$
	One day	1.00	1.00	1.34	2.22	835.0	.069	-0.191	$[-0.50, 4.32 \times 10^{-5}]$
	One week	1.50	1.00	1.70	2.66	1080.0	.70	-0.041	$[-4.57 \times 10^{-5}, 7.73 \times 10^{-5}]$
Six accounts	One hour	1.08	1.33	1.34	1.85	761.0	.025	-0.23	$[-0.33, -2.66 \times 10^{-5}]$
	One day	1.40	2.00	2.10	2.85	846.5	.13	-0.16	$[-0.80, 7.44 \times 10^{-5}]$
	One week	1.50	2.00	2.40	3.71	852.0	.15	-0.15	$[-1.00, 2.68 \times 10^{-5}]$

Table 5: Number of attempts tried per log-in task for successful log-in tasks. Log in after one hour, one day and one week corresponds to immediate, short-term and long-term log-in tasks, respectively. The Bonferroni-corrected threshold p-value is .0083. The result shows that participants from the two groups required a similar number of attempts to log into one account successfully.

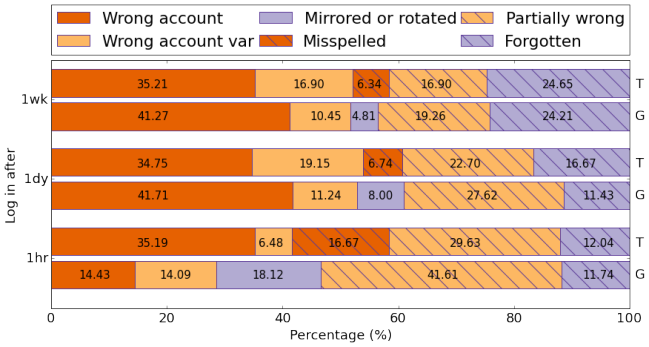


Figure 7: Categories of reason to fail at log-in tasks. Log in after one hour (1hr), one day (1dy) and one week (1wk) corresponds to immediate, short-term and long-term log-in tasks, respectively. Bars with label "T" are for text group and ones with "G" are for gesture group. In the legend, "wrong account var" stands for wrong account variant.

case when participants tried to log in one account with the password of another account, and "wrong account variant" was when participants used a password from another account but input incorrectly. "Partially wrong" indicated only part of the input matched the correct password. When participants tried random inputs, we categorized it as "forgotten". "Mirrored" and "rotated" categories was for gesture passwords: in both cases the input was correct but was either mirrored in direction, or rotated for a certain number of degrees. "Misspelled" was for text passwords only, which meant the input was correct except for one or two obvious typos.

We found nearly half of the errors both groups made were due to confusing one account with another ("wrong account" & "wrong account variant"). One exception is that one hour after creation, less than 30% of the errors made from the gesture group was due to confusing the accounts. Meanwhile, they made 10% more errors of partially incorrect inputs than text group. In the other two sessions, their partially-incorrect errors were similar. We note that errors with mirrored or rotated inputs for gesture group were nearly 20% after one hour, and decreased continuously thereafter.

Group	Duration (seconds) Mean (Median)	Duration (seconds) Max	# of attempts tried Mean (Median)	# of attempts tried Max
Text	67.19 (43.00)	450	5.42 (5)	21
Gesture	61.29 (39.00)	304	10.75 (7)	67

Table 7: Descriptive statistics for given-up log-in tasks. Duration is the average time participants spent on a single login task before they gave up, and the number of attempts is the retry rate.

On the other hand, two groups gave up a similar amount of tasks. 48 participants (52.75% of total) gave up on 192 log-in tasks in total. 78 of them (40.63%) occurred after one day, and another 98 (51.04%) were given up after a week. Text and gesture groups gave up 99 and 93 log-in tasks respectively.

A more detailed description of given-up tasks is in Table 7. The table shows that the gesture group spent less time but were willing to retry more times than the text group before they gave up. As the data of given-up log-ins was small, we examined it by combining data from different account settings and sessions. On average, our participants spent 64.31 seconds (Mdn=41.50) and tried eight times before giving up.

Subjective User Feedback

Exit Interview Questions

In the exit interview, we asked participants to estimate their daily usage of smartphones. The result shows the participants believed on average they entered passwords 8.93 times a day (SD=13, Mdn=4). Previous studies reported 8.11 times per day, but indicated it could be an underestimate [25]. Our study required at most twelve times a day.

Subjective Task Load Assessment

We asked participants to fill the NASA TLX form after every task. We calculated the average score per task per participant for two groups, with Figure 8 showing the result. A Wilcoxon rank sum test was applied on the TLX scores, and the test result showed there were no statistically significant difference in TLX scores between the two groups.

Account set	Log in after	Median (s) Text	Median (s) Gesture	Mean (s) Text	Mean (s) Gesture	Wilcoxon rank sum test W	p	r	95% C.I.
Two accounts	One hour	13.00	7.50	13.04	8.68	1632	$< .0001*$	-.4977	[2.83, 6.00]
	One day	11.25	7.00	12.43	8.45	1606	$< .0001*$	-.4762	[2.25, 5.25]
	One week	11.25	6.75	13.74	8.42	1559	$< .0001*$	-.4368	[2.00, 5.42]
Six accounts	One hour	9.92	6.86	11.64	7.69	1558	$< .0001*$	-.4358	[1.88, 4.98]
	One day	9.75	6.00	8.42	6.66	1656	$< .0001*$	-.5178	[2.40, 5.47]
	One week	8.54	5.75	8.64	6.25	1696	$< .0001*$	-.5511	[2.22, 4.65]

Table 6: The attempt duration of the two password groups in seconds. Log in after one hour, one day and one week corresponds to immediate, short-term and long-term log-in sessions, respectively. The Bonferroni-corrected threshold p-value is .0083. The result shows the attempt duration of the gesture group was much less than that of the text group in every login task.

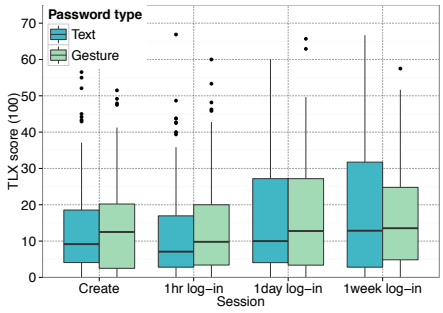


Figure 8: The TLX scores of two groups were similar across six individual factors. We calculated the average of the score of those factors.

DISCUSSION

We presented the first field study of free-form gesture passwords and compared them to traditional text passwords. We obtained a unique dataset by leveraging the Experience Sampling Method (ESM) for a password field study.

Usability

Our main finding in usability was that gesture passwords were faster in many aspects of authentication than text passwords. In our study, the gesture group took less time to both generate a new password and log in successfully when there were six accounts. On average, the gesture group spent 22% less time logging in, and 42% less time generating passwords than the text group. Also, performing a log-in attempt with a gesture required two to six seconds less than with text for most cases. This result matches intuition: drawing is faster than typing.

This is an important result for mobile platforms, given that the interactions are known to be fragmented and fast and as such may only last a few seconds [44, 28, 58, 62]. In this scenario, the authentication speed becomes pivotal for improving user experience; faster authentication allows for faster access to services on mobile devices while demanding less attention from the user during interaction. This speed advantage makes gesture passwords more suitable for mobile authentication.

This speed advantage does not necessarily come with the trade-off of complexity. We computed similarity score between gestures from same participant, and found that over 62% of participants had an average score below 2, indicating within-participant gesture reuse was limited. In addition, we found some gesture passwords were obviously complex and involve multiple fingers, words, signatures, and foreign language symbols. Still, the time people spent creating them was less than the average time amongst the gesture group.

We found gesture passwords were easier to use. Participants with gesture passwords were willing to retry as much as 46 more times than those with text passwords before they gave up on a log-in task. This might be because of the faster operation the former offered. Even with more retries, the time the gesture group spent on those tasks was six seconds less than that of the text group on average.

Participants found getting used to gesture passwords not trivial. There were several aspects showing that a learning curve existed for participants using gestures. First, the duration needed by the gesture group to create passwords or log-in decreased over time, while that of the text group either remained the same or increased. Second, a large portion of errors made by the gesture group shortly after the creation task was partially wrong, and such errors were reduced greatly over time. It is possible that at first participants were not familiar with the concept and our authenticator, therefore even if they knew the password they still failed. They essentially used such errors as training, and generated much better attempts for proceeding log-ins.

As a result, for a novel authentication scheme such as free-form gesture passwords, explicit practice sessions are desirable. During the introduction phase of our study, we suggested participants to practice their passwords during creation tasks until they feel comfortable, so long as the tasks have not expired. However, it is possible that in a field study as ours, where nobody was watching, participants did not put too much effort in practicing, but tried to complete tasks as quickly as possible. Consequently, they did not get familiar with gesture passwords before they proceeded.

Memorability

Gesture passwords provided good memorability with a log-in success rate over 83% after a week with six accounts. Considering the "learning curve" issue discussed above, this result demonstrated that gesture passwords were similarly memorable as text passwords.

We also discovered gesture passwords provide better distinction between multiple accounts: the gesture group was less confused with different accounts. In particular, participants with gesture passwords made 20% less "wrong account" errors than those with text passwords one hour after creation (see Figure 7). The text group, on the other hand, maintained a consistent percentage of such errors over time.

Gesture passwords had their own novel memorability issues. One hour after the creation process, participants tended to confuse the angle of their password. As a result, 18.12%

of their errors were due to mirroring or rotating of the correct passwords (see Figure 7). To compare, the text group made 16.67% of their errors at the same time interval due to mistyping. The portion of mirrored or rotated errors was even larger than that of the mistyping errors made by the text group. Understanding the novelty effect and reducing corresponding errors could be a key part in significantly improving the memorability of gesture-based authentication systems.

Security

We presented the analysis of the first free-form gesture password set collected in the field. Comparing with a study on shortcut gestures on mobile platforms [47], we found gesture passwords were different from shortcut gestures. The shortcut gestures study had half of the collected gestures as letters and only 10% were shapes. In contrast, in our field study dataset, nearly 50% were shapes. Both letter-shaped gesture passwords in our study and theirs were relatively simple; typically, the first letter of the account name (e.g. 'm' for music). This makes sense for shortcut gestures, but they are easy passwords to guess and repeat. As such, we postulate that our participants preferred shapes over letters for security reasons.

Interestingly, when comparing to a lab study where participants were also asked to create "secure and memorable gestures" [51], there were major differences in gesture creation: roughly half of the passwords generated in the previous study were with single finger, whereas 93% of our passwords were using one finger. There are three possible conjectures for this. First, the previous study did not involve multi-account interference as our study: each participant of that study generated only one password, while each of our participants generated eight. Second, it could be that people tend to overestimate the security of gesture passwords [51]. It is crucial to understand the gap between the security of novel authentication schemes and the perception of it from users. Third, it is possible that participants generated weaker passwords in a field study than lab study. Previous studies reported similar observations between a lab and a field study [22, 2, 53]. The text passwords created in our study were weaker than usual as well: most of them were easy, consisted of very few non-lowercase letters, and highly crackable (see Table 2).

Our proposed security metric also compared the entropy of our text and gesture passwords. For the grid of cells the screen split into, 4x6 is still considered "low resolution", but our analysis (Figure 6) showed it already contained similar entropy as the third quantile of our text passwords. This indicated that our gesture and text passwords were comparable in terms of security. Based on this metric and its model, we could also derive a naive guessing attack for gestures by generating points to fall into any of the cells of the screen, and connect the points to form a guess. Although cracking attacks of free-form gestures are beyond our scope, it could be an interesting future topic.

Completion Rate

The completion rate of our study was similar to that of a previous study on multi-password interference [21]. However, our participants completed more tasks per person (8 creation + 24 log-in), our expiration time was shorter (every task expired in one hour), and our exclusion rule was more strict (we only included participants who completed at least half of the designated tasks). To compare, the previous work mostly set the expiration time as one day or more [21, 62]. Our better completion rate is likely due to: (1) our expiration time was shorter, and (2) our notifications were native to the phone, and required no Internet or cellular access. Such a design lowered the effort for participants to complete tasks.

Limitations

In our experiment, we provided the same general instructions for generating passwords for both groups, which might lead to weaker passwords compared to specific policies. However, free-form gestures have no established composition policies so far, because it is a recently proposed approach. Also, not giving specific instructions allowed us to collect data on how people would use such a novel scheme in the wild. This data can then potentially shed light on how to design policies.

Our methodology is limited by common issues of a field study: lack of complete control over participants and the experiment. However, using ESM in the experiment design allowed us to have control over aspects such as task schedules and the amount of tasks each participant received. We believe our study maintained better control compared to conventional field studies while still collecting real-world data.

Our security metric is based on random entropy, which has been criticized for its bias [4]. Therefore, the result obtained should be interpreted only for relatively comparing security of the two, not measuring the absolute security of passwords.

CONCLUSIONS

Gesture-based interactions are becoming increasingly popular and prevalent on mobile platforms. We performed a study of free-form gesture passwords on mobile devices *in the field*, with text passwords as a baseline. We leveraged ESM to ensure the study was carried out under real-world settings with good control. Our study is the first step towards real-world usage of free-form gesture passwords. The study shows that free-form gesture passwords are faster to perform, faster to create, faster to log in, and have strong log-in success rates.

To conclude, the findings suggest that free-form gestures bear potential as an authentication method that is particularly suitable for mobile devices. We believe our observations and dataset are unique and present valuable data for designing secure and usable mobile authentication systems. Datasets and more material available at `http://securegestures.org`.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Number 1228777. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Gradeigh D. Clark was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

REFERENCES

1. Anne Adams and Martina Angela Sasse. 1999. Users Are Not the Enemy. *Commun. ACM* 42, 12 (Dec. 1999), 40–46. DOI: http://dx.doi.org/10.1145/322796.322806
2. Florian Alt, Stefan Schneegass, Alireza Sahami Shirazi, Mariam Hassib, and Andreas Bulling. 2015. Graphical Passwords in the Wild: Understanding How Users Choose Pictures and Passwords in Image-based Authentication Schemes. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '15)*. 316–322. DOI: http://dx.doi.org/10.1145/2785830.2785882
3. Lisa Feldman Barrett and Daniel J Barrett. 2001. An introduction to computerized experience sampling in psychology. *Social Science Computer Review* 19, 2 (2001), 175–185.
4. Joseph Bonneau. 2012. The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy (SP '12)*. IEEE Computer Society, Washington, DC, USA, 538–552. DOI: http://dx.doi.org/10.1109/SP.2012.49
5. Joseph Bonneau, Cormac Herley, Paul C. van Oorschot, and Frank Stajano. 2012. The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy (SP '12)*. IEEE Computer Society, 553–567. DOI: http://dx.doi.org/10.1109/SP.2012.44
6. Joseph Bonneau and Sören Preibusch. 2010. The Password Thicket: Technical and Market Failures in Human Authentication on the Web. In *9th Annual Workshop on the Economics of Information Security, WEIS 2010*. http://weis2010.econinfosec.org/papers/session3/weis2010_bonneau.pdf
7. Jon Brodkin. 2012. 10 (or so) of the worst passwords exposed by the LinkedIn hack. (2012). http://arstechnica.com/security/2012/06/10-or-so-of-the-worst-passwords-exposed-by-the-linkedin-hack/.
8. ChaosComputerClub. 2013. Chaos Computer Club breaks Apple TouchID. (2013). Retrieved May 3 2015 from http://www.ccc.de/en/updates/2013/ccc-breaks-apple-touchid.
9. Sonia Chiasson, Alain Forget, Elizabeth Stobert, P. C. van Oorschot, and Robert Biddle. 2009. Multiple Password Interference in Text Passwords and Click-based Graphical Passwords. In *Proceedings of the 16th ACM Conference on Computer and Communications Security (CCS '09)*. 500–511. DOI: http://dx.doi.org/10.1145/1653662.1653722
10. Karen Church, Mauro Cherubini, and Nuria Oliver. 2014. A Large-scale Study of Daily Information Needs Captured in Situ. *ACM Trans. Comput.-Hum. Interact.* (2014), 10:1–10:46. DOI: http://dx.doi.org/10.1145/2552193
11. Gradeigh D. Clark and Janne Lindqvist. 2015. Engineering Gesture-Based Authentication Systems. *Pervasive Computing, IEEE* (Jan 2015), 18–25. DOI: http://dx.doi.org/10.1109/MPRV.2015.6
12. Mihaly Csikszentmihalyi and Reed Larson. 1987. Validity and reliability of the experience-sampling method. *The Journal of nervous and mental disease* 175, 9 (1987), 526–536.
13. Mihaly Csikszentmihalyi, Reed Larson, and Suzanne Prescott. 1977. The ecology of adolescent activity and experience. *Journal of youth and adolescence* 6, 3 (1977), 281–294.
14. Alexander De Luca, Alina Hang, Frederik Brudy, Christian Lindner, and Heinrich Hussmann. 2012. Touch Me Once and I Know It's You!: Implicit Authentication Based on Touch Screen Patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. 987–996. DOI: http://dx.doi.org/10.1145/2207676.2208544
15. Alexander De Luca, Alina Hang, Emanuel von Zezschwitz, and Heinrich Hussmann. 2015. I Feel Like I'M Taking Selfies All Day!: Towards Understanding Biometric Authentication on Smartphones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. 1411–1414. DOI: http://dx.doi.org/10.1145/2702123.2702141
16. Alexander De Luca, Marian Harbach, Emanuel von Zezschwitz, Max-Emanuel Maurer, Bernhard Ewald Slawik, Heinrich Hussmann, and Matthew Smith. 2014. Now You See Me, Now You Don'T: Protecting Smartphone Authentication from Shoulder Surfers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. 2937–2946. DOI: http://dx.doi.org/10.1145/2556288.2557097
17. Alexander De Luca and Janne Lindqvist. 2015. Is secure and usable smartphone authentication asking too much? *Computer* 48, 5 (May 2015), 64–68. DOI: http://dx.doi.org/10.1109/MC.2015.134
18. Alexander De Luca, Emanuel von Zezschwitz, Ngo Dieu Huong Nguyen, Max-Emanuel Maurer, Elisa Rubegni, Marcello Paolo Scipioni, and Marc Langheinrich. 2013. Back-of-device Authentication on Smartphones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. 2389–2398. DOI: http://dx.doi.org/10.1145/2470654.2481330
19. Serge Egelman, Sakshi Jain, Rebecca S. Portnoff, Kerwell Liao, Sunny Consolvo, and David Wagner. 2014. Are You Ready to Lock?. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14)*. 750–761. DOI: http://dx.doi.org/10.1145/2660267.2660273

20. Serge Egelman, Andreas Sotirakopoulos, Ildar Muslukhov, Konstantin Beznosov, and Cormac Herley. 2013. Does My Password Go Up to Eleven?: The Impact of Password Meters on Password Selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. 2379–2388. DOI: http://dx.doi.org/10.1145/2470654.2481329

21. Katherine M. Everitt, Tanya Bragin, James Fogarty, and Tadayoshi Kohno. 2009. A Comprehensive Study of Frequency, Interference, and Training of Multiple Graphical Passwords. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. 889–898. DOI: http://dx.doi.org/10.1145/1518701.1518837

22. Sascha Fahl, Marian Harbach, Yasemin Acar, and Matthew Smith. 2013. On the Ecological Validity of a Password Study. In *Proceedings of the Ninth Symposium on Usable Privacy and Security (SOUPS '13)*. 13:1–13:13. DOI: http://dx.doi.org/10.1145/2501604.2501617

23. Andy Field, Jeremy Miles, and Zoë Field. 2012. *Discovering statistics using R*. SAGE Publications. 428–429 pages.

24. Edwin A Fleishman and James F Parker Jr. 1962. Factors in the retention and relearning of perceptual-motor skill. *Journal of Experimental Psychology* 64, 3 (1962), 215.

25. Dinei Florencio and Cormac Herley. 2007. A Large-scale Study of Web Password Habits. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*. 657–666. DOI: http://dx.doi.org/10.1145/1242572.1242661

26. Michael Frank, Ralf Biedert, En-Di Ma, Ivan Martinovic, and Dong Song. 2013. Touchalytics: On the Applicability of Touchscreen Input as a Behavioral Biometric for Continuous Authentication. *Information Forensics and Security, IEEE Transactions on* (2013), 136–148. DOI: http://dx.doi.org/10.1109/TIFS.2012.2225048

27. Google. 2013. Google NGram Viewer. (2013). Retrieved May 3 2015 from http://storage.googleapis.com/books/ngrams/books/datasetsv2.html.

28. Marian Harbach, Emanuel von Zezschwitz, Andreas Fichtner, Alexander De Luca, and Matthew Smith. 2014. It's a Hard Lock Life: A Field Study of Smartphone (Un)Locking Behavior and Risk Perception. In *Symposium On Usable Privacy and Security (SOUPS 2014)*. USENIX Association, Menlo Park, CA, 213–230. https://www.usenix.org/conference/soups2014/proceedings/presentation/harbach

29. Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology* (1988), 139–183.

30. hashcat. 2015. oclHashcat - advanced password recovery. (2015). Retrieved May 2 2015 from http://hashcat.net/oclhashcat/.

31. Eiji Hayashi and Jason Hong. 2011. A Diary Study of Password Usage in Daily Life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. 2627–2630. DOI: http://dx.doi.org/10.1145/1978942.1979326

32. Robin M. Hogarth, Mariona Portell, and Anna Cuxart. 2007. What Risks Do People Perceive in Everyday Life? A Perspective Gained from the Experience Sampling Method (ESM). *Risk Analysis* 27 (2007), 1427–1439. DOI: http://dx.doi.org/10.1111/j.1539-6924.2007.00978.x

33. KoreLogic Inc. 2015. KoreLogic Security. (2015). Retrieved May 2 2015 from http://www.korelogic.com.

34. Philip G. Inglesant and M. Angela Sasse. 2010. The True Cost of Unusable Password Policies: Password Use in the Wild. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. 383–392. DOI: http://dx.doi.org/10.1145/1753326.1753384

35. Ari Juels and Madhu Sudan. 2006. A fuzzy vault scheme. *Designs, Codes and Cryptography* 38, 2 (2006), 237–257.

36. KoreLogic. 2010. KoreLogic HashCat Rules. (2010). Retrieved May 1 2015 from http://contest-2010.korelogic.com/rules-hashcat.html.

37. John Lawler. 1999. The "web2" File of English Words. (1999). Retrieved May 3 2015 http://www-personal.umich.edu/~jlawler/wordlist.

38. Yang Li. 2010. Protractor: A Fast and Accurate Gesture Recognizer. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. 2169–2172. DOI: http://dx.doi.org/10.1145/1753326.1753654

39. Nicholas Micallef, Mike Just, Lynne Baillie, Martin Halvey, and Hilmi Güneş Kayacik. 2015. Why Aren'T Users Using Protection? Investigating the Usability of Smartphone Locking. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '15)*. 284–294. DOI: http://dx.doi.org/10.1145/2785830.2785835

40. Wendy Moncur and Grégory Leplâtre. 2007. Pictures at the ATM: Exploring the Usability of Multiple Graphical Passwords. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. 887–894. DOI: http://dx.doi.org/10.1145/1240624.1240758

41. Miguel A. Nacenta, Yemliha Kamber, Yizhou Qiang, and Per Ola Kristensson. 2013. Memorability of Pre-designed and User-defined Gesture Sets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. 1099–1108. DOI: http://dx.doi.org/10.1145/2470654.2466142

42. James Nicholson, Lynne Coventry, and Pam Briggs. 2013. Age-related Performance Issues for PIN and Face-based Authentication Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. 323–332. DOI: http://dx.doi.org/10.1145/2470654.2470701

43. Antti Oulasvirta, Anna Reichel, Wenbin Li, Yan Zhang, Myroslav Bachynskyi, Keith Vertanen, and Per Ola Kristensson. 2013. Improving Two-thumb Text Entry on Touchscreen Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. 2765–2774. DOI: http://dx.doi.org/10.1145/2470654.2481383

44. Antti Oulasvirta, Sakari Tamminen, Virpi Roto, and Jaana Kuorelahti. 2005. Interaction in 4-second Bursts: The Fragmented Nature of Attentional Resources in Mobile HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*. 919–928. DOI: http://dx.doi.org/10.1145/1054972.1055101

45. Allan Paivio and Kalman Csapo. 1973. Picture superiority in free recall: Imagery or dual coding? *Cognitive psychology* 5, 2 (1973), 176–206.

46. Martin Pielot, Karen Church, and Rodrigo de Oliveira. 2014. An In-situ Study of Mobile Phone Notifications. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services (MobileHCI '14)*. 233–242. DOI: http://dx.doi.org/10.1145/2628363.2628364

47. Benjamin Poppinga, Alireza Sahami Shirazi, Niels Henze, Wilko Heuten, and Susanne Boll. 2014. Understanding Shortcut Gestures on Mobile Touch Devices. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services (MobileHCI '14)*. 173–182. DOI: http://dx.doi.org/10.1145/2628363.2628378

48. Napa Sae-Bae, Kowsar Ahmed, Katherine Isbister, and Nasir Memon. 2012. Biometric-rich Gestures: A Novel Approach to Authentication on Multi-touch Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. 977–986. DOI: http://dx.doi.org/10.1145/2207676.2208543

49. M.Angela Sasse, Michelle Steves, Kat Krol, and Dana Chisnell. 2014. The Great Authentication Fatigue And How to Overcome It. In *Cross-Cultural Design*. Springer International Publishing, 228–239. DOI: http://dx.doi.org/10.1007/978-3-319-07308-8_23

50. Muhammad Shahzad, Alex X. Liu, and Arjmand Samuel. 2013. Secure Unlocking of Mobile Touch Screen Devices by Simple Gestures: You Can See It but You Can Not Do It. In *Proceedings of the 19th Annual International Conference on Mobile Computing & Networking (MobiCom '13)*. 39–50. DOI: http://dx.doi.org/10.1145/2500423.2500434

51. Michael Sherman, Gradeigh Clark, Yulong Yang, Shridatt Sugrim, Arttu Modig, Janne Lindqvist, Antti Oulasvirta, and Teemu Roos. 2014. User-generated Free-form Gestures for Authentication: Security and Memorability. In *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '14)*. 176–189. DOI: http://dx.doi.org/10.1145/2594368.2594375

52. Aaron Smith. 2015. U.S. Smartphone Use in 2015. (2015). Retrieved August 27 2015 from http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/.

53. Youngbae Song, Geumhwan Cho, Seongyeol Oh, Hyoungshick Kim, and Jun Ho Huh. 2015. On the Effectiveness of Pattern Lock Strength Meters: Measuring the Strength of Real World Pattern Locks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. 2343–2352. DOI: http://dx.doi.org/10.1145/2702123.2702365

54. Trustwave SpiderLabs. 2012. Hey, I just met you, and this is crazy, but here's my hashes, so hack me maybe? (2012). Retrieved Feb 12 2015 from https://www.trustwave.com/Resources/SpiderLabs-Blog/Hey,-I-just-met-you,-and-this-is-crazy,-but-here-s-my-hashes,-so-hack-me-maybe-/.

55. SRLab. 2013. Spoofing fingerprints. (2013). Retrieved Feb 12 2015 from https://srlabs.de/spoofing-fingerprints/.

56. Elizabeth Stobert and Robert Biddle. 2013. Memory Retrieval and Graphical Passwords. In *Proceedings of the Ninth Symposium on Usable Privacy and Security (SOUPS '13)*. Article 15, 14 pages. DOI: http://dx.doi.org/10.1145/2501604.2501619

57. Jing Tian, Chengzhang Qu, Wenyuan Xu, and Song Wang. 2013. KinWrite: Handwriting-Based Authentication Using Kinect. In *Proceedings of NDSS 2013*.

58. Sebastian Uellenbeck, Markus Dürmuth, Christopher Wolf, and Thorsten Holz. 2013. Quantifying the Security of Graphical Passwords: The Case of Android Unlock Patterns. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security (CCS '13)*. 161–172. DOI: http://dx.doi.org/10.1145/2508859.2516700

59. Blase Ur, Sean M. Segreti, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Saranga Komanduri, Darya Kurilova, Michelle L. Mazurek, William Melicher, and Richard Shay. 2015. Measuring Real-World Accuracies and Biases in Modeling Password Guessability. In *24th USENIX Security Symposium (USENIX Security 15)*. 463–481. https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/ur

60. Emanuel von Zezschwitz, Alexander De Luca, and Heinrich Hussmann. 2014. Honey, I Shrunk the Keys: Influences of Mobile Devices on Password Composition and Authentication Performance. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational (NordiCHI '14)*. 461–470. DOI: http://dx.doi.org/10.1145/2639189.2639218

61. Emanuel von Zezschwitz, Alexander De Luca, Philipp Janssen, and Heinrich Hussmann. 2015. Easy to Draw, but Hard to Trace?: On the Observability of Grid-based (Un)Lock Patterns. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. 2339–2342. DOI: http://dx.doi.org/10.1145/2702123.2702202

62. Emanuel von Zezschwitz, Paul Dunphy, and Alexander De Luca. 2013. Patterns in the Wild: A Field Study of the Usability of Pattern and Pin-based Authentication on Mobile Devices. In *Proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI '13)*. 261–270. DOI: http://dx.doi.org/10.1145/2493190.2493231

63. Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* (1945), 80–83.

64. Yulong Yang, Janne Lindqvist, and Antti Oulasvirta. 2014. Text Entry Method Affects Password Security. In *The LASER Workshop: Learning from Authoritative Security Experiment Results (LASER 2014)*. https://www.usenix.org/conference/laser2014/program/agenda/presentation/yang

65. Nan Zheng, Kun Bai, Hai Huang, and Haining Wang. 2014. You Are How You Touch: User Verification on Smartphones via Tapping Behaviors. In *Network Protocols (ICNP), 2014 IEEE 22nd International Conference on*. 221–232. DOI: http://dx.doi.org/10.1109/ICNP.2014.43