USABLE SECURITY: HUMAN FACTORS IN MOBILE AUTHENTICATION

by

YULONG YANG

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Electrical and Computer Engineering

Written under the direction of

Janne Lindqvist

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2016

This research is based upon work supported by the National Science Foundation under Grant Numbers 1228777, 1211079 and 1223977. Any opinions, findings, and conclusions or recommendations expressed in this research are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

ABSTRACT OF THE DISSERTATION

Usable security: human factors in mobile authentication

By YULONG YANG

Dissertation Director:

Janne Lindqvist

Text passwords are still the primary authentication mechanism for computers and online systems world-wide. Prior work indicates that they would likely persist in the foreseeable future, despite alternative proposals. Therefore, it is crucial to examine the open issues in text passwords. In addition, instead of replacing text passwords entirely, alternatives could be proposed for use under specific context. Under such premises, this thesis focused on (1) to demonstrate the field performance of a serious alternative method for mobile authentication and (2) to propose a systematic experiment design to study password memorability.

Designed to be used for desktop computers originally, text passwords are not suitable for modern platforms such as mobile devices. Using text passwords on mobile devices is a drastically different experience, because of the different form factor and context. From a between-group lab study comparing passwords usage on different devices, we learned that the form factor alone already has an effect on aspects of passwords such as the amount of lowercase letters used per password.

Meanwhile, recent studies suggest that free-form gesture passwords are a viable alternative as an authentication method on touchscreen devices. However, little is known

about the actual advantages they carry when deployed for everyday mobile use. We performed the first field study (N=91) of mobile authentication using free-form gestures, with text passwords being the baseline. Motivated by Experience Sampling Method (ESM), our study design aimed at increasing ecological validity while still 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 the choice of gestures varied across categories. Our findings demonstrated gesture passwords are a serious alternative for mobile context.

A major struggle people have with text passwords is to create ones that are both secure and memorable. Although there has been research on measuring password security, we have yet to systematically discover the factors to affect password memorability. By combining existing memory findings and password specific contexts, we proposed a field experiment design centering on two major factors that affect password memorability: log-in frequency and password condition. Log-in frequency defines the frequency of log-in tasks, and password condition defines the condition each password was created. The result of the experiment revealed that potential effects of our factors exist and pointed out directions for future studies.

Acknowledgements

I would not have been able to finish my Ph.D. thesis without the guidance and support from colleagues and those close to me.

First and foremost, I would like to express my sincere gratitude to my advisor, Professor Janne Lindqvist. When I started working with him, I had little prior experience in academic research or the field of human computer interaction. Janne guided me, from scratch, towards my first experiment protocol, first user study, first paper and first publication. He was dedicated and committed on educating and directing me throughout my Ph.D. journey. He was also there to motivate me and push me beyond my limit, I literately would not have been able to complete the work I have done so far without him.

I am fortunate enough to have collaborated with many brilliant minds, which I deeply appreciated. Professor Antti Oulasvirta is one of the kindest person I have ever met. We have collaborated on different projects since the beginning of my research. Although most of the time we communicated through emails from different time zones, Antti offered me concrete help and insightful advices on how to proceed in my research. Thanks to Can Liu and Xianyi Gao, who helped me conduct user studies for the memorability project, I was able to concentrate more on writing this thesis. Gradeigh D. Clark and I have worked on several projects together, and he provided his unique insight and excellent writing skills. I would also like to thank my other collaborators and co-authors, including Michael Sherman, Ben Firner, Huiqing Fu, Shridatt Sugrim and Professor Yang Wang.

One of the things I appreciated the most is that I worked in WINLAB at Rutgers. It is such a big unique community with people coming from so many different places. I was lucky enough to get the chance to work and hang out with them.

I would also like to thank Professor Wade Trappe, Professor Richard Martin and Dr. Robert Miller for their time serving on my committee.

Most of my friends have little idea of what kind of research I am working on, and they do not need to know. One of the most incredible things I had during the years of Ph.D. is their friendship. We share experiences and emotions in life with each other, good or bad. The life of an international student far from home is not always the best experience. Having someone to hang out with or talk to makes everything better.

Special thanks to my girlfriend, Yanran Wang. Our relationship made a big difference and warmed me during my Ph.D. years. I appreciate everything we have together from past to present, especially for the times when she had to change herself for me, which I know is very difficult for any person to do.

Finally, I would like to thank my parents, who always support and care about me. We have arguments now and then, but I know how remarkable and difficult it is for them to come such a long way and understand my opinions and decisions. In the end, their unconditional support makes me tough enough to face everything.

Dedication

To the future of passwords.

Table of Contents

Abstract ii

Acknowledgements iv

Dedication vi

List of Tables xi

List of Figures xiv

1. INTRODUCTION 1

1.1. Overview 1

1.2. Organization 3

1.3. Contribution 3

2. BACKGROUND AND RELATED WORK 5

2.1. Mobile Authentication 5

2.1.1. Experience Sampling Method in Password Study 8

2.2. Password Security and Usability 9

2.3. Password Memorability 11

2.3.1. Existing Memory Theories 13

3. TEXT ENTRY METHOD AFFECTS PASSWORD SECURITY . . 14

3.1. Overview of Chapter 14

3.2. Method 15

3.2.1. Experiment Design 15

3.2.2. Apparatus 16

3.2.3. Procedure 17

3.2.4. Participants . . . 18
3.2.5. Password Security Estimation . . . 19
3.2.6. Password Cracking Attacks . . . 20
3.3. Results . . . 21
3.3.1. Structures . . . 21
3.3.2. Quantitative Password Security . . . 24
3.3.3. Cracking Attacks . . . 24
3.3.4. Task Load . . . 28
3.4. Discussion . . . 29
3.5. Summary . . . 31

4. FREE-FORM GESTURE AUTHENTICATION IN THE WILD . . 32
4.1. Overview of Chapter . . . 32
4.2. Method . . . 32
4.2.1. Participants . . . 33
4.2.2. Experiment Design . . . 33
Experience Sampling Method . . . 34
4.2.3. Apparatus . . . 35
Gesture Password Authenticator . . . 37
4.2.4. Procedure . . . 37
4.2.5. Password Analysis . . . 39
Security Comparison . . . 39
4.2.6. Statistical Tests . . . 40
4.3. Results . . . 40
4.3.1. Creation Tasks . . . 40
Duration . . . 40
Text Passwords Created . . . 41
Free-form Gesture Passwords Created . . . 43
Security Comparison . . . 44

4.3.2. Log-in Tasks . . . 45
Log-in Success Rate . . . 45
Duration . . . 45
Attempts . . . 46
Errors . . . 47
4.3.3. Subjective User Feedback . . . 50
Exit Interview Questions . . . 50
Subjective Task Load Assessment . . . 50
4.4. Discussion . . . 51
4.4.1. Usability . . . 51
4.4.2. Memorability . . . 53
4.4.3. Security . . . 53
4.4.4. Completion Rate . . . 54
Limitations . . . 55
4.5. Summary . . . 55

5. FACTORS IN PASSWORD MEMORABILITY . . . 57
5.1. Overview of Chapter . . . 57
5.2. Method . . . 58
5.2.1. Experiment Design . . . 58
Controlled Variable . . . 58
Password Strength Meter . . . 60
Tasks . . . 62
Schedule . . . 63
Proactive Interference . . . 63
5.2.2. Apparatus . . . 64
5.2.3. Procedure . . . 65
Revisions Based On Pilot Study . . . 65
5.2.4. Analysis . . . 66

5.2.5. Participants . . . 67
5.3. Results . . . 67
5.3.1. Pilot Study . . . 67
5.3.2. Formal Study . . . 70
Creation . . . 71
Log-in . . . 74
Exit Survey . . . 74
Factors To Affect Memorability . . . 76
5.4. Discussion . . . 78
5.4.1. Limitations . . . 78
5.4.2. Pilot Study . . . 78
5.4.3. Formal Study . . . 79
5.5. Summary . . . 81

6. CONCLUSIONS . . . 82

References . . . 83

Appendix A. Memorability study . . . 94
A.1. Application Screenshots . . . 94
A.2. Accounts . . . 94

List of Tables

3.1. Definition of each category of passwords. All types with low occurrence in our passwords were aggregated into "others" category. 23

3.2. Results of both plain dictionary attacks and long-session offline attacks. "Include" listed all dictionaries we used in each attack. The size was the number of unique entries each combined dictionary had for dictionary attacks, and the number of guesses generated per password for long-session offline attacks. Facebook attack performed the best on Smartphone group, and Password attack worked best on Laptop and Tablet group compared with Words and Facebook attacks. It suggested passwords of different groups carried different level of resistance against cracking attacks. 26

4.1. 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. The completion rate was high compared with previous studies. 41

4.2. Text passwords' number of characters (left), and cracking attack result (right).The cracked result is similar to that of the weakest category of passwords being cracked under the similar experiment setup. 43

4.3. 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. The results show that success rate of two groups was mostly similar across conditions. 46

4.4. 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. 47

4.5. 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. 48

4.6. 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. 49

4.7. 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. In general gesture group spent less time while were willing to retry more in given-up log-in tasks. . . . 49

5.1. Password strength meter setup for the two password conditions. The naive check disallowed the exact same password across multiple accounts, and the similarity check set a threshold for password reuse based on edit distance. 61

5.2. General password statistics, including password length, amount of lowercase letters, uppercase letters, digits and symbols per password. The results show that passwords generated in our study are non-trivial ones with complex structure. 71

5.3. Table of Pearson's correlation coefficient r and corresponding p value computed between each per-account factor and the log-in success rate. Factors such as num of symbols are the amount of characters in a single password. Factors such as log-in attempts are the average value of each account. Factors with prefix "zxcvbn" were computed by the zxcvbn password strength meter. As indicated, attempts and duration factors show strong correlation with log-in success rate. 77

5.4. Table of Pearson's correlation coefficient r and corresponding p value computed between each per-task factor and the log-in status of the task. "No. of the task" refers to the order of this task. "Account" differentiates accounts from each other. Attempts, duration, the order of the task and log-in frequency show strong correlation with log-in status. 78

A.1. The account information of our memorability study. For some accounts we used the description to help participants better understand the type of the acocunt. 94

List of Figures

3.1. The keyboard layout for the devices in the tablet group and smartphone group. Note that two groups shared the same key positions within each layout, but the structures of the four layouts were different for each: the tablet group followed the more common structure, while the smartphone group had a hierarchical structure. To reach the next layout of the smartphone keyboard, one had to first reach the previous one. Therefore, the smartphone keyboard had a higher difficulty reaching non-lowercase keys than the tablet keyboard. 17

3.2. The average password length, amount of lowercase letters, uppercase letters, digits and symbols appeared in single password across groups. Error bars stand for 95% confidence intervals based on a bootstrap (that is, not assuming normality). The figure shows notable difference in password length and amount of lowercase letters across groups. 22

3.3. A comparison of distribution of passwords in different categories for each group. The most-common category was different across groups, indicating passwords generated by different text entry methods have different resistance against cracking attacks. 24

3.4. The mean score of three password security metrics across groups: score from the Adaptive Password-Strength Meter (APSM), random entropy and NIST entropy. Error bars stand for 95% confidence intervals based on a bootstrap (that is, not assuming normality). The figure shows that passwords from different groups share similar estimate by all three security measurements. 25

3.5. The percentage of passwords cracked by our first offline attack. The x-axis was in log10 scale. The final percentage of cracked passwords for laptop group, tablet group and smartphone group were 14.2%, 17.3% and 15.6%, respectively. The cracking results across groups were similar. 27

3.6. The percentage of passwords cracked by Weir's algorithm vs. the number of guess, per group. The x-axis was in log10 scale. The final percentage of cracked passwords for laptop group, tablet group and smartphone group are 31.7%, 28.4% and 30%, respectively. The cracking results across groups were similar. 27

3.7. A comparison of percentages of cracked passwords in different categories across groups. The percentage value shows the percentage of cracked password in total amount of passwords in each group. We kept the categories and percentage scale as the same as in Figure 3.3 for better comparison. Cracked passwords here were the combination of cracked passwords in all our attacks. The most-cracked category was different for each group, according to the figure, indicating the different resistance each group of passwords possessed. 28

3.8. Mean score for each item in TLX form in session one and two. Error bars stand for 95% confidence intervals based on a bootstrap (that is, not assuming normality). The figure shows three groups have similar ratings across all individual factors. 29

4.1. 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. 36

4.2. 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. 38

4.3. 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 figure shows the gesture group has an overall shorter duration than the text group in most categories. 42

4.4. 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. Based on the figure, "Shapes" and "Letters" were favored by participants. 44

4.5. 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. Within "Shapes" category, participants preferred common shapes. 44

4.6. 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. The figure shows that modeling the touchscreen to be 3x5 grid of cells matched the median entropy of our text password. 45

4.7. 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. The most notable thing is that the gesture group made fewer errors of "Wrong account" (and variant) than the text group in log-in tasks after one hour. 50

4.8. The TLX scores of two groups were similar across six individual factors. We calculated the average of the score of those factors. The ratings were similar between groups. 51

5.1. The plot of log-in success rate of different log-in frequency (a) and password condition (b). Both two factors show a visible effect on log-in success rate. In both figures blue lines indicate the success rate, with the Y-axis being on the left; the red boxplots indicate log-in duration data, with the Y-axis being on the right. For duration boxplots, the black squares indicate the mean value. According to the figure, the log-in success rate decreased when the log-in frequency became lower, or the password condition changed from simple to strong. 69

5.2. Log-in success rate of the first log-in task for all accounts grouped by frequency (retention effect, left figure), and log-in success rate of all the log-in tasks in order as repetitions (practice effect, right figure). Neither curves show an obvious pattern, indicating no retention effect or practice effect in our data. 69

5.3. The boxplot of creation duration for two password conditions. Red squares indicate the mean value. It shows that participants spent more time in generating passwords under strong condition than the simple condition. 71

5.4. The boxplot of password length and characters count for two password conditions. The characteristics of passwords in two conditions are not very different from each other. 72

5.5. The result of our cracking attacks. X-axis represents different cracking methods and their total guesses, and each line represents a password dataset. “Field study” is from chapter four, “TextEntry” is from chapter three. “Pilot” and “Formal” are from this chapter. “Formal+name” used the same dataset as “Formal”, but adding variations of account names in our study to the input dictionary to create a targeted attack, as mentioned in the method section. The figure shows that datasets from this study have better resistance against cracking than others, and our targeted attack cracked nearly 10% more of the “Formal” dataset than the regular attack. 73

5.6. Figure (a) is the log-in success rate and duration of different log-in frequency; Figure (b) is that for different password conditions. In both figures blue lines indicate success rate, with Y-axis being on the left; red boxplots indicate log-in duration data, with Y-axis being on the right. For duration boxplots, the black squares indicate the mean value. The figure shows a decreasing trend in success rate when log-in frequency lowers, but no visible effect of password condition on the rate. 75

5.7. Log-in success rate of the first log-in task for all accounts grouped by frequency (retention effect, left figure), and log-in success rate of all the log-in tasks in order as repetitions (practice effect, right figure). The retention effect in the left figure is very subtle, and we can see a visible practice effect in the right figure. 75

A.1. A screenshot of our web application. It shows the creation interface for one of the accounts. 95

A.2. A screenshot of our web application. It shows the log-in interface for one of the accounts. 95

Chapter 1

INTRODUCTION

1.1 Overview

Text-based passwords remain as the most prevalent method of authentication [69]. In addition to traditional computers such as desktops and laptops, people increasingly generate and use passwords with a wide variety of mobile terminals, such as tablets and smartphones. They also store a large amount of personal and sensitive data (e.g. banking information, home address) [42, 121].

Because interactions on mobile terminals are drastically different from the traditional computers [99], it is crucial to understand whether such differences affect generated passwords in a similar way. One of such difference is the *text entry method* [87], which consists of those physical (e.g. form factor, display) and software (e.g. virtual keyboard layout) aspects of an input device that are relevant when entering text. The design of a text entry method determines how quickly and effortlessly a given character can be typed. Even small changes in how characters are displayed and organized can affect typing performance [141]. Furthermore, one should see corresponding differences in the distribution of characters in different methods. For instance, digits are not directly reachable without changing the layout in the common touchscreen qwerty keyboard on smartphones – does this affect the generated passwords?

In addition to the text passwords, alternative authentication methods have been proposed and deployed on mobile platforms, such as PIN, grid-pattern lock and biometrics. However, they suffer from various shortcomings: limited password space [128], susceptibility to shoulder surfing [38, 133], easily crackable [17, 23, 124], slow entry [37] and potentially harmful to privacy [35].

Free-form gesture passwords have been recently proposed as an alternative for mobile user authentication [29,118]. Free-form gesture passwords allow users to draw any shape or pattern on a blank touchscreen display using one or more fingers. Previous studies demonstrated that, for both mobile and non-mobile use, free-form gesture passwords can be secure and memorable [118, 127]. They potentially provide a large password space because of their free-form nature. Gesture-based interaction conforms to the form factor of mobile devices [104] and is faster than typing. Since mobile interactions tend to be fragmented, frequent and short-term [62, 101], gesture passwords could be more suitable for authentication on mobile devices. In addition, when used as a password, free-form gestures improve memorability with the help of visual learning effects [103] and motor memory [49].

Most work on gesture passwords so far has been carried out in laboratories [36, 54, 105, 118], 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 [36, 54, 105, 113, 118], and has not considered it for multi-account configurations. However, people manage multiple accounts at the same time in their daily lives [50, 68]. Previous work showed multi-account settings affected the authentication process: a study showed that multi-account interference significantly impacts the ease of authentication of facial graphical passwords [46]. Therefore, it is crucial to explore how gesture passwords would be different under the multi-account context.

One of the most mentioned issues of text passwords is the difficulty of generating secure yet memorable passwords. People utilize many strategies to avoid forgetting passwords: password reuse, frequently resetting passwords, or end up generating weak passwords. However, comparing with the extensive research on password security, that of password memorability remains largely unexplored. Studies have used memorability as a metric to evaluate authentication schemes or strategies [25, 27, 45, 73, 91, 140], but seldom examined how exactly each component of such schemes or strategies influences

the memorability. Here, one of the research challenges is to properly design the experiment to include potential factors from different aspects, such as password-related (password length, password security), usage-related (log-in frequency, account type) and behavior-related (password reuse).

We believe it is important to quantitatively understand, in the process of using a password, what could be the factors to affect its memorability. Such knowledge could enable us to design authentication schemes with a good balance of security and memorability. Therefore, as the third part of the thesis, we present an experiment design that aimed at exploring the memorability of passwords systematically. We then used two studies to obtain initial results and further refine the design.

1.2 Organization

Chapter 2 describes background and related work on password security, usability, memorability, and mobile authentication research. Chapter 3 presents the laboratory study in which we explored the effect of text entry methods on passwords. Chapter 4 describes the field study of mobile authentication where we discovered similarity and difference of text passwords and free-form gesture passwords. Chapter 5 presents our methodology to study password memorability. We then conclude the thesis in Chapter 6.

1.3 Contribution

This thesis presents the following contributions:

1. We provided insight into passwords generated by different text entry methods, as well as how people recalled them. We found that, although the effect of text entry methods was not as significant as we hypothesized, it did exist in our study: more lowercase letters were used in passwords on mobile devices.

2. This thesis presented the first field study on *memorability* and *usability* of free-form gesture passwords. We recruited 91 participants who generated 708 passwords across two weeks and recalled them at three different points of time.

3. We found that free-form gesture passwords are more resilient to multi-account interference than text passwords, and provide better mobile usability than text passwords. Our analysis of the first field gesture dataset showed that the choice of gesture passwords by participants were varied.

4. This thesis proposed the first systematic experiment design for password memorability. The design focused on two major factors: log-in frequency and password condition. Our studies found log-in frequency has strong effect on memorability while that of password condition is limited. We also identified the effect of other password-specific factors such as password reuse and password characteristics.

Chapter 2

BACKGROUND AND RELATED WORK

In this chapter, we review previous literature that motivated our work. Many design choices in our experiments were inspired from them. Reviews are in the topics of mobile authentication, and three vital aspects of passwords: security, usability, memorability.

2.1 Mobile Authentication

Many researchers have studied the usability of mobile platforms for passwords. Greene et al. [59] studied the difference between typing passwords using tablets and smartphones in a between-group experiment. They found that the time it took participants to type and recall passwords significantly varied depending on the source of entry methods. The time was also different given different passwords. Schaub et al. [107] found similar significant time differences among different smartphones. In addition, they found that attackers had significantly different success rates in shoulder surfing passwords on those smartphones. Both of the mentioned studies did not ask participants to create passwords, but provided participants with passwords instead, thusly having little information on password generation and consequently how password security would be affected if created using different entry methods.

Few studies have specifically looked at helping people to create passwords on mobile text entry methods. Haque et al. [60] have studied how to create secure passwords on mobile devices. They found the entropy of passwords were significantly different across mobile keyboards. However, only an approximation of Shannon entropy was examined. An analysis with additional security metrics and password structures could help us gain more insight into the effects of text entry methods. Jakobsson et al. proposed fastwords,

which relied on standard error-correcting features for users to create passphrases [73]. It was designed for non-traditional devices such as mobile handsets, and offered advantages of speed of entry, and good recall rate.

Recently, Bonneau et al. [13] studied how people chose 4-digit PINs for banking accounts. Common strategies included birth dates and visual patterns. The reported presence of visual strategies supported our hypothesis that passwords generated with different text entry methods, too, may differ.

Free-form gesture passwords have been proposed as a mobile authentication method and evaluated to be secure and memorable in the lab. Sherman et al. developed an information-theoretic metric to estimate the security of free-form gesture passwords. The metric was motivated by a study on information capacity of continuous full-body movements [100]. They conducted a lab study showing that single-finger gestures and gestures with many hard angles and turns achieved better security and memorability [118]. Another lab study, focusing on user interfaces and not authentication, found that user-defined gestures demonstrated better memorability than pre-defined ones [92]. Previous study also indicated that free-form gestures were resistant against a major threat on mobile platforms: shoulder surfing [118].

There are also studies that combine gestures with other factors for authentication: simple strokes on the mobile device itself [36] and tapping actions [143]. Biometric-based gesture authentication has been proposed to utilize the unique way each person performs an identical set of template gestures [105]. Another work extracted features of users when they performed simple and day-to-day tasks on smartphones such as scrolling or swiping, and used that to verify users [54]. Although, previous work showed that brute-force attacks with input data from the general population was able to compromise biometric-based authentication systems [112]. Another study showed that users lack originality when generating gestures for HCI tasks [95]. They observed that users often repeat known gestures or use common ones. In addition to mobile device gestures, researchers have explored mid-air gestures for authentication [127].

Other alternatives have been proposed as well. A grid-based graphical password has been proposed for touchscreen devices to overcome the input accuracy issues on such

platforms [25]. It allows users to continuously draw their passwords by "warping" on multiple layers, enabling them to generate more complex passwords on a small screen. A detailed review of the various types of graphical passwords was written by Biddle et al. [8]. Biometric authentication (e.g. keystroke-based [20] and touch-based [34]) has also been proposed for mobile platforms.

Finally, there are studies on smartphone unlocking behaviors. Android grid pattern has been deployed to most modern Android smartphones, where users draw a secret by connecting dots in a typically 3x3 grid. A study looked at the motivations of why some users choose to (or not to) employ locking mechanisms for their devices [44]. They reported a strong correlation between the use of locking and users' risk perceptions, and the likely underestimation of the privacy risks of users. Due to the design of fixed-position grids, the password space of grid pattern is limited. A large-scale study (N=105) estimated security of Android Pattern Unlock based on Markov chains [128]. They discovered that user-chosen patterns were biased to a few pattern selection strategies. As a result, the estimated security in entropy was less than that of 3-digit randomly-assigned PINs. Another study claimed that it is possible to retrieve the partial or even complete version of some Android grid patterns based solely on the smudge traces on the touchscreens [4].

Apple deployed *Touch ID*, a fingerprint-based biometric authentication scheme. Three user studies were carried out to demonstrate that Touch ID did not help users create stronger passcodes to lock their phones [24]. They also found a mismatch between the expectations from participants towards the security of their passcodes and the reality. Another work used an online survey to discover that usability is the top reason why people adopt biometric authentication methods such as Touch ID and FaceUnlock in Android [35]. Surprisingly, privacy risk was seldom mentioned by participants when asked why they not adopt such measures. The work also raised the necessity for biometric authentication methods to be "socially compatible" in design. Recently, a gesture-unlocking scheme was proposed. It utilized user input features such as finger velocity, device acceleration and stroke time [113].

2.1.1 Experience Sampling Method in Password Study

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 [46]. 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 [45]. 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 [134]. Another study focused on designing and implementing strength meters for pattern unlock and found that it improved the security [122]. 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 [125]. Schneegass et al. used both a lab study and a field study to introduce *SmudgeSafe*, an authentication system based on random image transformations for unlocking touchscreen devices [109]. Their system used transformed images (flipped, scaled, rotated) to prevent attackers to reconstruct correct password based on smudge traces on the touchscreen.

There is, however, limited literature applying ESM to mobile authentication studies. One study utilized ESM to capture participants' perceptions towards unlocking behaviors, revealing the reasonings behind leaving a phone unlocked [62]. 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 [72]. The study found that existing password policies are often beyond the capabilities of people who used them. Another study asked participants to record password events when they log into their accounts using desktop computers or laptops [68]. A diary study showed that authentication tasks lowered the productivity of employees in an organization [106].

2.2 Password Security and Usability

There has been extensive research in the security and usability of passwords.

Shannon entropy [114] has been used to measure password security [19], and has been criticized by more recent literature [10, 137]. Several more recent security metrics have been proposed. Bonneau et al. proposed a partial guessing metric to estimate the security of large-scale password distribution [10]. Instead of estimating the guesses required to crack the entire dataset, the proposed metric, α-guesswork, focused on the guesses required to crack part of it. Other work reported the password security estimate based on different cracking algorithms [39, 77]. They assumed the threat model of text passwords to be long-session offline attacks. They computed the number of guesses various cracking algorithms needed to crack each password, and used that as a security measurement. Castellucia et al. proposed an adaptive password strength meter that computed the entropy of a password based on the N-gram and Markov model [22].

Password strength meters have been wildly deployed in the industry to help users generate passwords. However, previous work examining meters adopted by major web sites on the Internet showed that most of those vendors did not provide any rationale for their design choices [21]. One exception was the *zxcvbn* password meter made by DropBox, which not only explained its design and logic [41], but also open-sourced the meter [40]. Similarly, a study examining various password manager softwares and built-ins found that their designs and policies were significantly different from each other, and some of which could be easily exploited [119]. Another work states that password meters lead to stronger passwords only for important accounts [45]. Ur et al. [129] found that stringently rated password meters led users to make significantly longer passwords that included more types of characters, and passwords were also more resistant against cracking algorithms.

Researchers have also studied how password generation policies have affected password security and usability. Weir et al. [137] claimed that passwords created under common requirements, such as minimum length and different character set requirements, were still vulnerable to cracking attacks. Shay et al. [117] found that some policies

that required longer passwords provided better usability and security compared with traditional policies. At the same time, a study based on online attacks of web accounts claimed that it is misguided to require users to generate strong passwords [51]. They argued that the combined size of username and password space should be considered in terms of security rather than password space alone. Schechter et al. proposed a novel password policy system that allows any passwords so long as they are not common ones [108]. They defined and populated common passwords by calculating their occurrences using count-min sketch, a bloom-filter-like algorithm to ensure the upper bound on the number of times a password appears in the dataset. Forget et al. proposed a system to place randomly-selected characters into arbitrary positions of the user-generated passwords, in the purpose of improving password security [53]. Although their system was found to significantly improve password security, they observed that participants intentionally chose weak passwords before the randomize mechanism to compensate memory load. Another study concluded that system-assigned passphrases did not seem to offer superior performance over system-assigned passwords, in terms of usability [116]. Zhang et al. designed an efficient framework to search the old password based on the new password of the same person [142]. They concluded that password expiration policies are inefficient and fail to meet the designed goal.

Other studies demonstrated that semantic and linguistic patterns affected security and usability as well. Bonneau et al. found that the choice of phrases for passphrases is not random [15]. Their results showed that users strongly preferred simple two-word phrases commonly used in English. Veras et al. demonstrated that there existed semantic patterns in user-chosen passwords, and one could exploit such patterns to boost the cracking performance [131]. A study comparing passwords generated by English and Chinese users found they were different from each other in many aspects [86]: Chinese passwords contained Pinyins and preferred digits, while English passwords contained English words and preferred lowercase letters. Their Chinese-specific algorithm increased the cracking efficiency by 34%.

Gaw et al. reported on average 7.8 accounts per undergraduate student using a combined approach. They first provided a list of web services for participants to choose

from, and then asked participants to recall additional items on themselves [55]. They also found that the majority of participants had only three or fewer unique passwords. A 3-month study obtained password-usage data by a browser plug-in [50]. It showed that people on average managed seven unique passwords, each of which were used for about 5.67 different sites. They noted that possible over-counting existed in reported numbers, and estimated on average 25 accounts per person using a browser client. A 2-week diary study estimated participants had an average of 11.4 accounts per person [68]. Grawemeyer [58] conducted a diary study and found people had different generation strategies for different accounts. Such studies indicated that multi-account scenario would be reasonable in password experiments.

Password re-use has become a common strategy in password management. A study showed that people categorized their passwords into a limited set of categories, with varied security. Accounts in higher categories were more important, like financial accounts [61]. They also found it possible to crack passwords from higher categories if that of lower categories were known, as they were similar to each other. An interview study revealed that people tend to use weaker passwords for most services even when they possessed longer passwords [132]. Florêncio et al. stated that password management strategies that rule out weak passwords and re-use are sub-optimal [52]. Another study suggested that a maximum of four or five passwords per person reaches the limit of most users' memory capabilities [1].

Finally, Fahl et al. [47] compared real passwords to those generated in an experiment, finding that about 30% of subjects did not behave as the same as in daily life. However, the authors concluded that laboratory studies generally create useful data.

2.3 Password Memorability

In previous studies, different metrics have been used to demonstrate password memorability. One commonly-used metric is the log-in success rate, which is commonly defined as the ratio of successful login tasks over the number of total login tasks for each participant, or a certain condition or group [9, 14, 25–28, 45, 70, 91]. One issue with

log-in success rate is that it is possible for all groups in an experiment to have a very high and similar success rate, making the comparison among groups difficult. Some previous studies avoided this by limiting the log-in attempts [26, 27]. Other metrics include successful log-in duration [125] and log-in attempts [93].

A set of factors has been discovered to pose effect on password memorability. Several studies revealed that repetitions helped people memorize passwords. By asking participants to memorize a secret gradually and repeatedly, a study claimed that 88% of its participants were able to recall a 56-bit secret code after three days [14]. A median of 36 log-ins was performed by its participants. Another study also utilized spaced repetitions to help participants memorize Person-Action-Object (PAO) stories, a password management scheme that helps people generate strong passwords [9]. 77% of their participants recalled all of their stories more than four months later, with at most 12 tests needed over that period. The study also found that the majority of forgetting occurred within the first 12 hours after the generation process. Similar finding were also stated in another study, in which they found that recalling after a short period of delay is an effective way to help retention [135].

In addition to repetitions, the frequency of such repetitions also affects password memorability. A study using a diary study and interviews reported that participants seldom forgot their passwords if the passwords were used frequently [72].

An analysis of system logs from 386 users found that less password recovery would have occurred if the authentication system allowed 10 retries instead of three [18], indicating log-in attempts affect memorability as well.

Password memorability was also found to be dependent on the number of accounts and passwords a person has. Studies showed that the number of accounts affects the memorability of many different password types including text passwords [27, 46, 135].

An online study found that chunking improved memorability of system-generated PINs [70]. Chunking refers to the method to break a single number into multiple shorter numbers for easier memorization.

Memorability has also been used as a metric to compare different password types. A study showed that graphical passwords resulted in a better log-in success rate than

text passwords in the short term, although the rate of two types were similar in the long term [27]. 4-digit PINs and graphical passwords were compared as well, with findings indicating that graphical passwords were more memorable [91]. Another study found that passwords based on mnemonic phrases were as easy to remember as naively selected passwords [140]. They also stated that passwords based on mnemonic phrases are as strong as random passwords in terms of resistance against cracking attacks. Studies that proposed novel authentication methods often evaluated the memorability of the proposed methods [25, 73].

2.3.1 Existing Memory Theories

The study of memory for verbal materials, like word lists, has been the topic of research for more than a century [79]. The typical experimental task involves studying materials and recalling them after a period of intervening activity. This process is called *learning*, or *relearning* if participants have already learned the same material before. Reading the series audibly once is considered as one trial. One metric to measure the memory performance is *savings*, which is defined as the ratio of saved time or number of trials in relearning compared with the learning process [43]. There have been several major findings in the memory function of learning process.

The *retention effect* looks at how much of the material people forget (and retain) over time. It was discovered that the forgetting occurred rapidly soon after the learning, and slowed down as the time went on [43]. Two theories have been proposed to explain the retention and forgetting. *Decay theory* states that forgetting is only due to the passage of time [126]. *Interference theory* argues that forgetting is caused by new events or materials occur between practice and test of the old material [89].

The *practice effect* looks at how practices of the material for subsequent relearning could be reduced after previous learnings. The results showed that despite the constant retention interval between each two relearnings, the number of practices required for each individual decreased steadily. Both the retention and practice effect follow a rapidly-decreasing function, which has been modeled using either the logarithm [43] or power function [3].

Other proposed effects include the *spacing effect*, which states that the performance of memory is better when the practices of materials to memorize is distributed over a certain period of time, instead of repeatedly learning in a short span of time. *Proactive interference* indicates that items learned earlier can interfere with items learned later [76]. It has also been found that given a list of items to learn, the most recent and least recent items could be recalled better than items in the middle of the list. Such an effect is called the *serial-position effect* [76].

Chapter 3

TEXT ENTRY METHOD AFFECTS PASSWORD SECURITY

3.1 Overview of Chapter

In this chapter, we examine whether the design of text entry methods affect the security of generated passwords. We hypothesize that, depending on the password generation strategy, users may generate passwords using the characters on the display as generation cues. More precisely, the difficulty to reach a character from the present layout should affect the probability of its inclusion in a password. This could manifest both password structure and password security. Therefore, we aim at discovering possible differences in both password structure and security.

We first examine whether the structure of generated passwords are different across text entry methods. Metrics of password structure include password length, the amount of lowercase letters, uppercase letters, symbols and digits per password. We will also look at the types of passwords. We define the stype of a password by the types of characters it contains.

We then explore the question of whether text entry methods affect password security. We estimate the security of passwords from two aspects: quantitative estimation and practical cracking attacks. Quantitative estimation included Shannon entropy [88, 114], NIST entropy [19] and a recently introduced Markov-model-based metric (adaptive password-strength meter [22]). Then we look at how resistant these passwords are against cracking attacks.

Finally, we study if participants perceive the task with different text entry methods differently using the NASA Task Load Index assessment (TLX) [63].

3.2 Method

Our study was conducted in a laboratory to control for confounding factors. A controlled laboratory experiment allowed for choosing the main factor to be considered, in our case the text entry method. Next, we describe our method in details.

3.2.1 Experiment Design

The experiment followed a between-group design with text entry method types (3 levels) as an independent variable. We divided our participants into three groups based on the text entry method they used. The participants were randomly assigned into one of these three groups, and were unaware of the assignments or that other groups existed. A detailed explanation of the differences between the groups is given in the next subsection.

The main reason for us to choose between-group was to isolate its effect from any other undesired effects such as any possible confounding factors that would correlate with both the variable and the result. We noticed that previous work that involved the password generation process also had a similar experiment design [27, 59, 107].

An alternative design would be within-subject, in which one participant would perform the same task using three different text entry methods in a sequence. In such a design, the use of different text entry methods would generate undesired interference among each other for each participant. In particular, learning and using one text entry method would interfere with the learning and using of other text entry methods, thus decreasing or even eliminating the potential effect of both methods. Such interference is common in paired tasks [6, 16].

Florencio et al. revealed that people manage multiple passwords in reality [50]. To increase ecological validity, we asked participants to manage three different virtual accounts. However, since the difference within each participant was not in our research objective, we did not analyze the difference among three passwords created by each participant. Instead, the mean value of three accounts was taken to represent each participant in our models.

3.2.2 Apparatus

Our text entry method variable was defined by the apparatus each group used.

Laptop group (control group)

We provided a common laptop (Macbook Pro 2012 with a 13" display) in the laptop group. We chose so because the physical laptop keyboard was still the most common text entry method for password creation.

Tablet group

We provided a Samsung Nexus 10 tablet (Android 4.2.2, 10.1" touchscreen) as the device used in tablet group. The touchscreen keyboard on the tablet had a common qwerty layout, as shown in Figure 3.1. Given that the tablet can be held in the hands in two ways, we asked the participants to keep it in the "landscape" mode.

Smartphone group

We provided a Samsung Galaxy Nexus (Android 4.2.2, 4.5" touchscreen) as the device used in the smartphone group. The keyboard layout was chosen from several available designs for smartphone platforms.

The difference between our smartphone keyboard and tablet keyboard was the number of key presses needed to reach certain keys (see Figure 3.1). To reach uppercase letters, one needed to press two additional keys from the first layout in a smartphone group, while only one key press in tablet group. Also, to reach special symbols layout, three additional key presses were needed in smartphone group while only two in tablet group. In short, reaching certain keys and switching layouts demanded more effort with the smartphone keyboard than with the tablet keyboard. The primary reason we chose our text entry methods so was to differentiate each group in their difficulty of reaching keys during text entry. All three apparatus still provided common usability for the particular platform.

Software

The application was implemented in both Python (for laptop group) and Java for Android (for smartphone and tablet group). It has two main features: password creation and password recall. In the password creation interface, participants were asked to

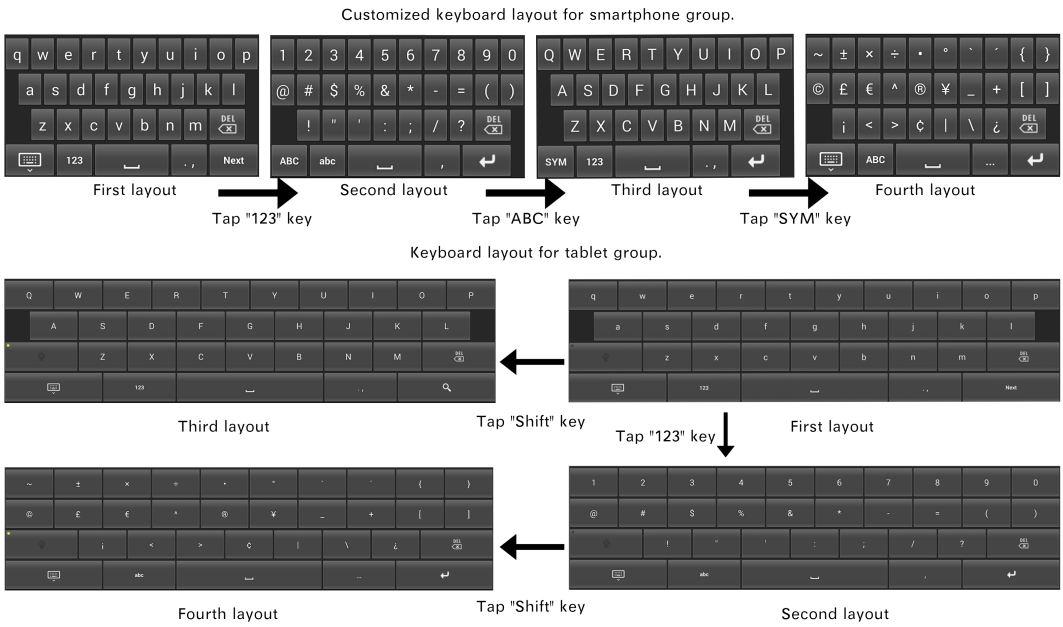


Figure 3.1: The keyboard layout for the devices in the tablet group and smartphone group. Note that two groups shared the same key positions within each layout, but the structures of the four layouts were different for each: the tablet group followed the more common structure, while the smartphone group had a hierarchical structure. To reach the next layout of the smartphone keyboard, one had to first reach the previous one. Therefore, the smartphone keyboard had a higher difficulty reaching non-lowercase keys than the tablet keyboard.

create usernames and passwords for three virtual accounts in the same order. Each virtual account had a different logo, color and short description. In the recall interface, it asked participants to recall what they created earlier for each account, in a different order. “Give up” button would show up after four failed attempts for each account.

3.2.3 Procedure

All experiments were conducted in the same office room we setup for this study.

The primary task for participants was to create and recall username and password for three different types of virtual accounts. We minimized the risk for participants by advising them not to use their existing passwords, and also keeping data in a safe place.

Our study consisted of two sessions. In session one, we asked participants to create a username and a secure password for three different accounts given a certain text entry

method. The detailed procedure was as follows:

1. **Introduction to the Study**. The participants were introduced to the study, which included reading and signing the consent form, discussion of their rights and also compensation.

2. **Password Creation**. Each participant was given the corresponding text entry method before the session. They were asked to create usernames and passwords for three different virtual accounts: bank, email and online magazine. The order of the accounts was the same for all participants at this step.

3. **Subjective Workload Assessment** The participants were asked to fill out the NASA TLX form [63].

4. **Distraction**. The participants were asked to do a mental rotation task [74] and count down from 20 to 0 mentally.

5. **Password Recall**. Participants were asked to recall usernames and passwords they created in the Password Creation step above. The order of the accounts were changed with Latin square. For each account, participants were allowed to try as many time as they wanted, and give up if necessary (showed up after four failed attempts).

6. **Survey**. Participants were asked several questions about password generation and also usual demographic questions.

In session two of our study, which was at least 10 days after session one, participants were asked to come back to recall the usernames and passwords. The recall procedure was the same as that in session one. After the recall process, participants were asked to fill out NASA TLX form and answer a few questions. We included recall sessions so as to avoid participants creating unrealistic passwords if they knew they would not need to recall such passwords afterwards.

3.2.4 Participants

We recruited participants through fliers, mailing lists, and in person at cafeterias. Participants were required to be over 18 years old and familiar with touchscreen devices. We recruited 63 participants in total, between the ages of 18 to 65 ($M = 27.2, SD = 9.9$). 24 of our participants were male and 39 were female.

All 63 participants completed session one of our study, and 57 of them returned for session two. As compensation, participants received one $30 gift card each for completing the whole study. They also participated in a raffle of three $75 gift cards.

We recruited our participants in two batches, 33 in May and 30 during June and July 2013. The gap between two sessions of the study varied. The mean time gap for the first batch was 14.53 (SD = 5.81) days and 29.52 (SD = 7.57) days for the second. The number of participants for the laptop group, tablet group and smartphone group were 21, 27 and 15, respectively.

Non-equal group sizes are expected after random assignment [110]. The tests applied in the following sections were applied to the entire sample distribution. To ensure the validity of results, we randomly sampled our two larger groups so that the size was even across groups, and then performed the same tests again. The results on the sampled data were the same, indicating our tests were robust against the unbalanced group size.

Our study was approved by the Institutional Review Board of Rutgers University.

3.2.5 Password Security Estimation

We describe our password security estimation below.

The first metric we used is Hartley entropy [65]. The entropy is defined in equation $H = L \times \log_2 N$, in which L is the length of the password, and N is the possible set of characters. Hartley entropy assumes the probability of every character got chosen to be in password is the same.

The NIST entropy was a scheme to evaluate human-selected passwords introduced in the NIST Electronic Authentication Guideline [19]. The scheme took into account the fact that passwords were chosen by human beings, who tend to choose passwords

that were easily guessed, and even from a set of a few thousand commonly chosen passwords. We implemented the scheme by assigning different entropy to characters at different positions, each password creation rule contributing to a specific amount of entropy and that the entropy of the policy was the sum of the entropy contributed by each rule. In addition, we performed a simple dictionary word check ("dic-0294") to give the password extra entropy.

The adaptive password-strength meter (APSM) based on Markov models estimated the strength of a password by estimating the probability of n-grams that composed the password [22]. N-gram is a contiguous sequence of n characters from a given string. Probabilities of n-grams are computed based on a large password dataset, therefore, it introduces certain dependency on the training password dataset. In our implementation, we used the "Rockyou" password dataset to compute the database of probabilities for every n-gram. The dataset contained over 32 million real passwords. We chose 4-gram as the element in our implementation as the original paper did.

There were some other metrics we did not include in our analysis. Bonneau has proposed several statistical metrics for password security [11]. However, Bonneau's metrics were mainly applicable to a large-scale password dataset, while we had a much smaller one.

3.2.6 Password Cracking Attacks

We performed several actual cracking attacks against our passwords. We used two popular password cracking tools, John the Ripper [96] and hashcat [66].

Dictionaries

We used various dictionaries that are common in the literature. "dic-0294" is a English dictionary from outpost9 [102]. "All" is a free public dictionary from openwall website [98]. "Mangled" is a paid dictionary from openwall. It is a hand-tuned wordlist containing four million password candidates generated using various mangle rules. "Rockyou" includes about 32 million passwords leaked from the website RockYou. "Facebook" is a list of names of searchable user from the website Facebook [120].

"Myspace" contains passwords from a phishing attack against MySpace website. "Inflection" [111] is a list of words along with their different grammatical forms such as plurals and past tense.

Our dictionary set included several password databases that were compromised and disclosed to public by hackers. While they are publicly available, we are aware of the fact that they contained sensitive information. We treated them confidentially, and disallowed any unauthorized access. Further, the security community in general had accepted several papers using such datasets, and thus seemed to consider it as an appropriate method.

Dictionary attack

First, we applied plain dictionary attacks using combinations of dictionaries. The first attack with "Words", which contained common words from different languages, aimed at easy passwords; the second with "Facebook", contained the entire directory from the website, aimed at passwords made with actual names, and popular phrases; the third attack with "Passwords", which contained common passwords and real leaked passwords, aimed at common and naive passwords.

Long session offline attack

We applied two long session attacks, simulating one attack with common resources and one longer attack with optimal strategies and more resources, respectively.

The first attack involved generating guesses based on a modified "Single mode" rules, which was originally from John the Ripper, using the "dic-0294" dictionary as input. The "Single mode" rules contained a set of rules to modify words including login names and directories to generate guesses [97]. The modified version, made by Weir [136], was optimized for the English dictionary. We followed the same setup of Weir et al. [137].

The second attack applied the probability password crack tool developed by Weir et al. [137,138]. It generated password guesses in the order determined by various rules derived from training sets. We used a similar model from experiment P4 conducted by Kelley et al. [78].

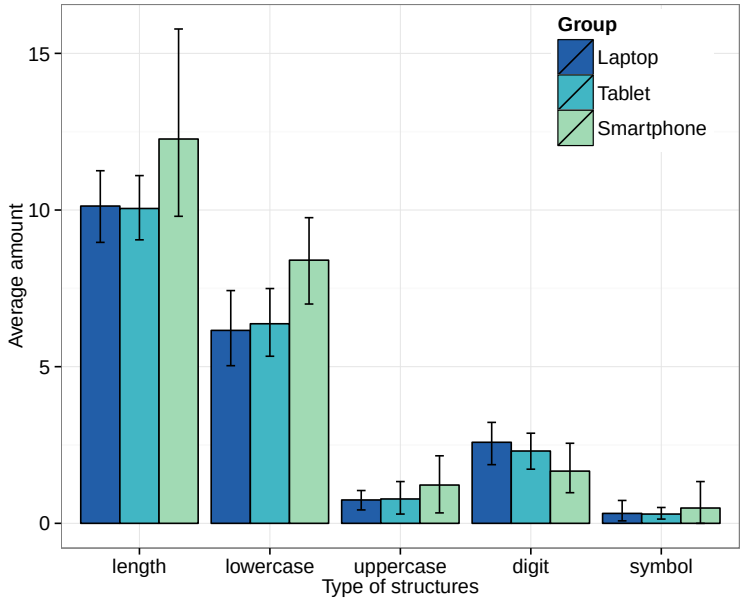


Figure 3.2: The average password length, amount of lowercase letters, uppercase letters, digits and symbols appeared in single password across groups. Error bars stand for 95% confidence intervals based on a bootstrap (that is, not assuming normality). The figure shows notable difference in password length and amount of lowercase letters across groups.

3.3 Results

We collected 189 passwords in total. In the following section we present our analysis results. The results focused on the analysis of password generation and password security, analysis of the passwords memorability is not included below.

3.3.1 Structures

Figure 3.2 shows the password length and the amount of characters per password classified by types across groups. It demonstrates a notable difference in password length and amount of lowercase letters between the smartphone group and other two groups.

For each structure metric, we performed an one-way ANOVA test across three groups. The text entry method variable had significant effect on the amount of lowercase letters, $F(2, 60) = 3.186$, $p = .048$, $\eta_p^2 = .066$. No significant results were found from other metrics.

Category	Description
loweralpha-num	only contains lowercase letters and digits
loweralpha	only contains lowercase letters
mixedalpha-num	contains lowercase and uppercase letters and digits
loweralpha-special-num	contains lowercase letters, special symbols and digits
all	contains lowercase and uppercase letters, special symbols and digits
mixedalpha	only contains lowercase and uppercase letters
others	types other than mentioned ones

Table 3.1: Definition of each category of passwords. All types with low occurrence in our passwords were aggregated into "others" category.

Next, we examined the categories of passwords each group generated. We defined the category of a password by types of characters it contained. The category of a password revealed the complexity in its structures: passwords containing multiple types of characters had a more complex structure than ones with only one type. Table 3.1 summarizes our definition of categories.

Figure 3.3 shows the distribution of passwords within the defined categories across groups. For smartphone group, passwords that contained only lowercase letters (*loweralpha*) was most common (31.1%). For other two groups, passwords containing only lowercase letters and digits (*loweralpha-num*) were the most common: 30.2% in laptop group and 38.2% in tablet group, respectively. In addition, there was no passwords containing lowercase letters, special symbols and digits (*loweralpha-special-num*) in smartphone group at all, while both other groups generated passwords in that category.

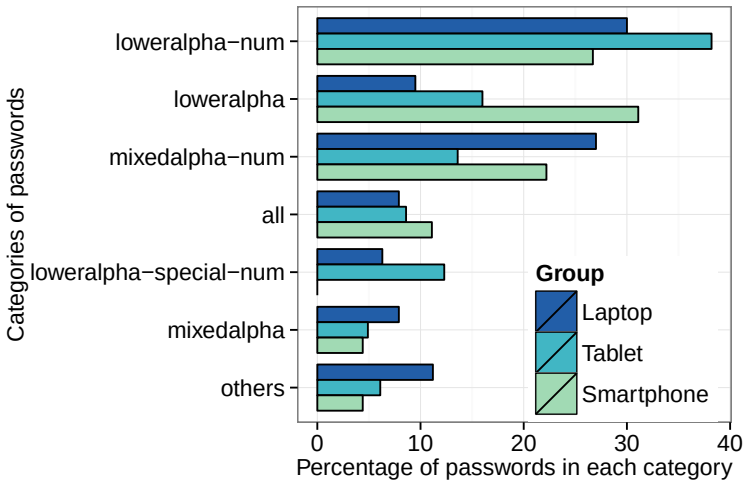


Figure 3.3: A comparison of distribution of passwords in different categories for each group. The most-common category was different across groups, indicating passwords generated by different text entry methods have different resistance against cracking attacks.

3.3.2 Quantitative Password Security

We estimated the security of our passwords with two common entropy-based password security metrics, random entropy and NIST entropy, and a more recent Markov model based metric (APSM). Such metrics provided quantitative measurement of password security. These metrics were explained in details in section 4.3. The mean scores and corresponding confidence intervals of the result are shown in Figure 3.4. According to the graph, scores of passwords from the smartphone group were consistently higher than that of other two groups. However, most of means stayed within the confidence interval of the value of other groups, indicating the differences among groups were limited.

We performed one-way ANOVA on the three sets of security measures. However, the results showed a non-significant effect of text entry method variable on them.

3.3.3 Cracking Attacks

We performed dictionary attacks and long-session offline attacks on our collected passwords. Both attacks have been described in details in section 4.4. Table 3.2 shows the result of plain dictionary attacks. The performance of “Words” and “Facebook” attacks were limited across all groups, except for the “Facebook” attack on passwords in

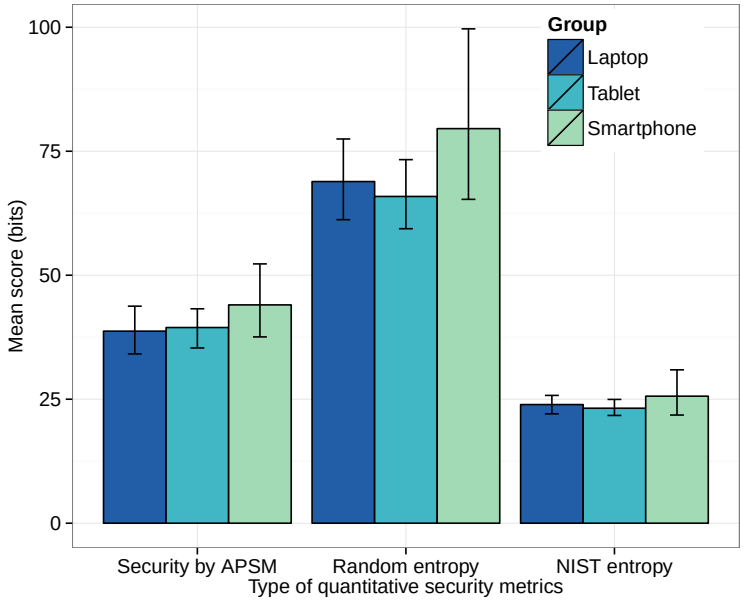


Figure 3.4: The mean score of three password security metrics across groups: score from the Adaptive Password-Strength Meter (APSM), random entropy and NIST entropy. Error bars stand for 95% confidence intervals based on a bootstrap (that is, not assuming normality). The figure shows that passwords from different groups share similar estimate by all three security measurements.

Name	Include	Size	Laptop (63)	Tablet (81)	Smartphone (45)
Words	"dic-0294", "all", "inflection"	4.1M	4 (6.3%)	4 (4.9%)	4 (8.9%)
Facebook	"facebook"	37.3M	3 (4.8%)	6 (7.4%)	7 (15.6%)
Passwords	"mangled", "rockyou"	54.8M	15 (23.8%)	12 (14.8%)	8 (17.8%)
Long-session 1	NA	1000M	9(14.2%)	14(17.3%)	7(15.6%)
Long-session 2	NA	20000M	20(31.7%)	23(28.4%)	13(30%)

Table 3.2: Results of both plain dictionary attacks and long-session offline attacks. "Include" listed all dictionaries we used in each attack. The size was the number of unique entries each combined dictionary had for dictionary attacks, and the number of guesses generated per password for long-session offline attacks. Facebook attack performed the best on Smartphone group, and Password attack worked best on Laptop and Tablet group compared with Words and Facebook attacks. It suggested passwords of different groups carried different level of resistance against cracking attacks.

smartphone group. The "Password" attack worked much better compared to the first two attacks against laptop and tablet group, while it had very limited improvement over previous attacks against smartphone group.

Figure 3.5 and Figure 3.6 show the results of two long-session offline attacks. According to the figures, although the lower bound of resistance (the number of guesses of the first cracked password) were different, the percentages of cracked passwords across groups were similar to each other.

When we combined cracked passwords from all the attacks together, the total number of cracked passwords for the laptop group, tablet group and smartphone group were 24 (38.1%), 24 (29.6%) and 16 (35.6%), respectively. Chi-square test had been performed on the cracked password ratio across groups, but no significant result was found ($\chi^2(2) = 1.21, p = 0.54$).

Figure 3.7 shows the distribution of all cracked passwords into different categories across groups, in which we saw quite different distributions. Particularly, the category with the largest percentage of cracked passwords was different for all three groups: *mixedalpha-num* (passwords contain uppercase letters, lowercase letters and digits) (10, 15.9%), *loweralpha-num* (13, 16.0%) and *loweralpha* (7, 15.6%), respectively.

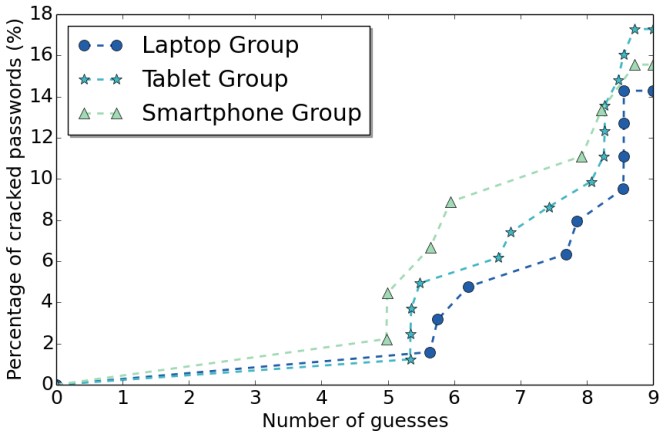


Figure 3.5: The percentage of passwords cracked by our first offline attack. The x-axis was in log10 scale. The final percentage of cracked passwords for laptop group, tablet group and smartphone group were 14.2%, 17.3% and 15.6%, respectively. The cracking results across groups were similar.

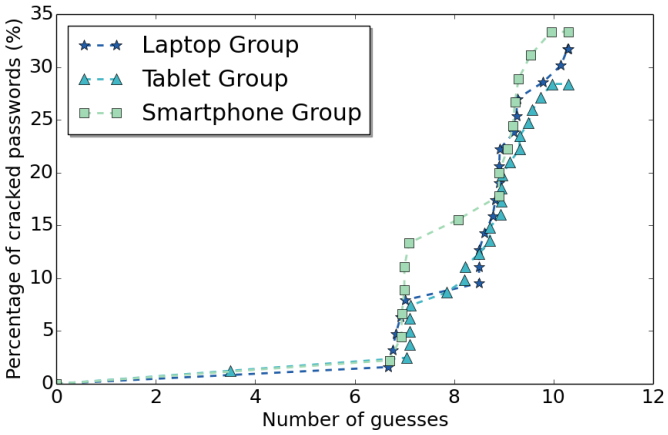


Figure 3.6: The percentage of passwords cracked by Weir's algorithm vs. the number of guess, per group. The x-axis was in log10 scale. The final percentage of cracked passwords for laptop group, tablet group and smartphone group are 31.7%, 28.4% and 30%, respectively. The cracking results across groups were similar.

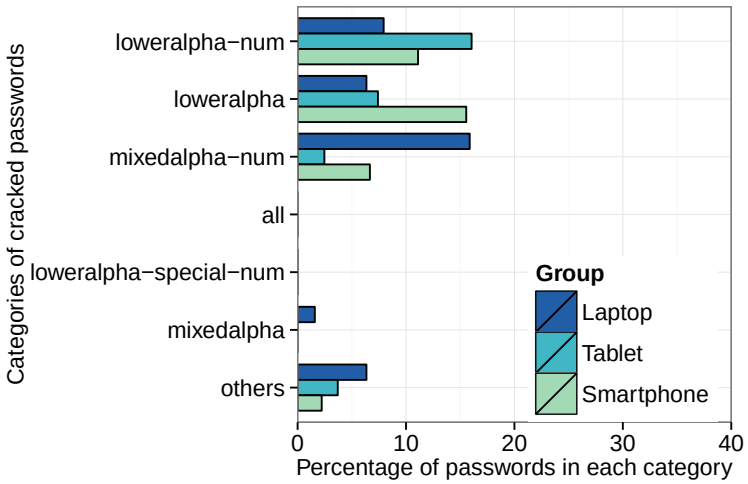


Figure 3.7: A comparison of percentages of cracked passwords in different categories across groups. The percentage value shows the percentage of cracked password in total amount of passwords in each group. We kept the categories and percentage scale as the same as in Figure 3.3 for better comparison. Cracked passwords here were the combination of cracked passwords in all our attacks. The most-cracked category was different for each group, according to the figure, indicating the different resistance each group of passwords possessed.

3.3.4 Task Load

We used TLX forms to evaluate the subjective task load of our study. These questions revealed participants' subjective assessment towards tasks in the study. Figure 3.8 shows the mean scores for each question of TLX form for both sessions.

Given individual items in one TLX form were correlated, we applied the MANOVA test with the text entry method as variable on the six items together, for session one and two, respectively. The result showed a non-significant effect of text entry method type on the scores of TLX assessment both for session one, $V = 0.21, F(8, 116) = 1.70, p = 0.11$, and session two, $V = 0.28, F(12, 100) = 1.37, p = 0.19$. Therefore, we concluded that participants in groups did not feel significantly different about the subjective task load of the experiment they participated in.

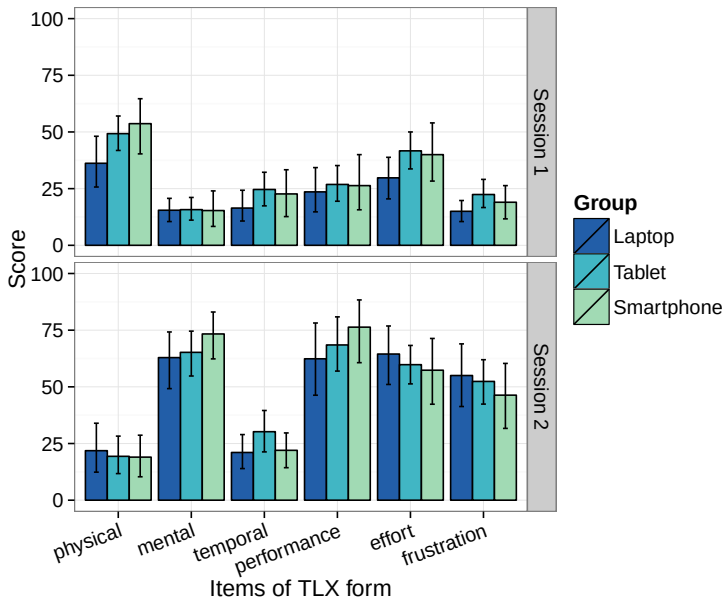


Figure 3.8: Mean score for each item in TLX form in session one and two. Error bars stand for 95% confidence intervals based on a bootstrap (that is, not assuming normality). The figure shows three groups have similar ratings across all individual factors.

3.4 Discussion

Our experiment successfully identified significant effect in password structures. In particular, passwords generated by the smartphone group consisted much more of lowercase letters per password than the other groups. However, quantitative security estimations, including random entropy, NIST entropy and score of APSM, did not differ significantly for passwords from the different groups.

One possible reason for this result could be that while passwords consisted of more lowercase letters were considered weaker, the smartphone group actually generated the longest passwords on average (around 12.5, compared to 10 in the other groups, see Figure 3.2). Extra length made passwords more secure. For example, a 15-character-long lowercase-only password from the smartphone group scored 101, 28.5 and 48 in random entropy, NIST entropy and APSM, respectively. All of which are well above the overall average.

In our study, the smartphone keyboard demanded the most effort in switching the layouts. As a result, participants switched layouts less often in the smartphone group,

leading to more lowercase letters in passwords. However, participants still placed sufficient effort on creating passwords, resulting in long passwords. According to Shay et al. [117], long passwords were generally more secure. Meanwhile, smartphone group participants did not report a higher load in TLX forms (Figure 3.8).

This is not to deny the fact that the difficulty in reaching non-lowercase letters affected password security for smartphone group. For two 10-character passwords from our study, the one with lowercase, uppercase and digits scored much higher than the one with only lowercase letters in our security estimation.

Therefore, one simple design modification for text entry methods in the smartphone group could be including digits or some special symbols in the first layout of the keyboard, without sacrificing usability. Such a design could encourage people to choose non-lowercase characters more often.

Also, the study is conducted as a lab environment, in which participants created passwords under the watch of experimenters. It is possible that under such conditions, participants spent extra effort to create passwords that are stronger than usual. For example, the average password length of each group in our study is at least 10 characters, while that of RockYou passwords is below 8.

In addition, whether the quantitative metrics we used reflected the true security of passwords is still a question. Random entropy and NIST entropy have been criticized in such a task [11, 137], which led us to include one more recent metic (APSM). We found that APSM could also compute quite different scores for very similar passwords. For example, "vowelword" and "bonesjones" were both lowercase-only letters consisting of two English words; however, APSM computed their scores to be 50 bits and 30 bits, respectively. This could be because APSM is dictionary dependent. Considering the mean score of APSM of our passwords were only 40 bits, a difference of 20 bits would be undismissible. Therefore, our study raised the need of a truly comprehensive and appropriate metric for gauging text password security.

On the other hand, the analysis of password structure and cracking attacks still showed the effect existed. As mentioned before, the variable had significant effect on number of lowercase letters in passwords (Figure 3.2). This finding was consistent

with our experiment design, as the difficulty of reaching non-lowercase keys in the smartphone group was increased. In addition, we found that passwords cracked in our attacks are distributed quite differently in categories across groups (Figure 3.7). Particularly, nearly 50% of cracked passwords in every group belonged to a different single category compared with each other. Such results indicated different resistance against cracking attacks across groups.

Limitations.

Our sample size was relatively small, a large-scale study would be desirable in the future. In addition, our study limited participants to create and recall passwords in a lab environment, which is not representative of the real scenario when passwords are used. While a recent study by Fahl et al. [47] showed that laboratory studies generally create useful data, a field study could be a follow-up on this topic. Also, while in our study we used common text entry methods, one could include more manipulations to see how would the effect be changed due to specific manipulations.

3.5 Summary

This chapter presented a randomized controlled laboratory experiment following a generate-test-retest experiment design. Our experiment discovered significant effect of text entry methods on the distribution of passwords created from them, and also their resistance against actual cracking attacks. It also confirmed prior results of how entropy-based metrics are not adequate to measure password security, including a more recent approach based on Markov models.

Chapter 4

FREE-FORM GESTURE AUTHENTICATION IN THE WILD

4.1 Overview of Chapter

In this chapter, we report the first field study using free-form gestures as a mobile authentication method, with text passwords as a baseline. The method is inspired by the Experience Sampling Method (ESM) [32], and includes two password-specific contexts: multi-account interference and variable recall time.

The chapter first explains our experiment design and justifications to support our decisions in the design. We then look at the creation tasks, examining the duration participants need to generate new passwords. We also analyze the first field gesture password dataset to understand the rationale when people generate their gesture passwords. Next we look at log-in tasks, comparing two authentication schemes under various contexts with several different metrics, including log-in success rate, duration, attempts and errors made in failed log-in tasks.

4.2 Method

Our study design follows established practices in ESM studies, using a smartphone as both the signaling and task performing device. The experiment followed a mixed design with between-group and within-subject variables. In this section, we describe our participants, experimental design, apparatus, and procedure.

4.2.1 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 [46, 91, 134]. 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 data in the format of 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.

4.2.2 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 [27, 46, 93]. 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 [50]. 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 [75].

Experience Sampling Method

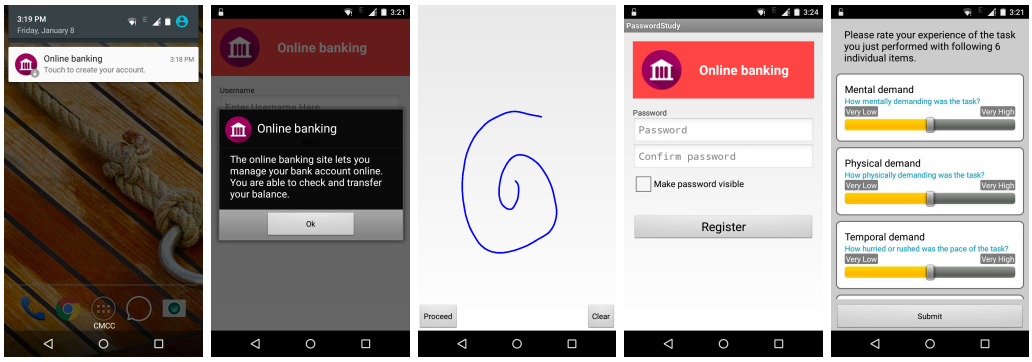
We incorporated the Experience Sampling Method (ESM) in our field study. ESM is a research methodology where participants perform tasks at different points of time during their daily lives. It is usually used to capture the subjective experience of participants in a natural environment [7, 31, 62].

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 [31], 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 [44, 50, 62, 90, 122]. 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 [7]. 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.

4.2.3 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 4.1 shows sample screen-shots of the application.



(a) Notification example

(b) Account description interface

(c) Gesture password input interface

(d) Text password input interface

(e) TLX form

Figure 4.1: 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.

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 4.1a shows a sample screen-shot of our task notification.

Tapping the notification brought participants to the user interface for either a password generation or log-in task. 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 [64] to fill out. The form is designed to estimate workloads subjectively using six individual factors [64]. 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 [34, 46, 125, 134]. 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

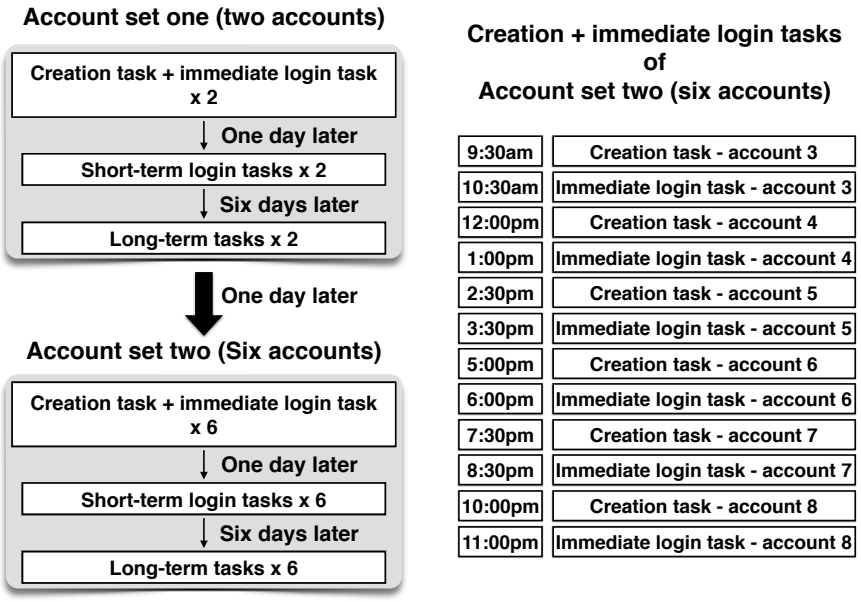
Our gesture password authenticator is a modification of Protractor [85] that was extended [118] for multi-touch gesture support. Protractor is a gesture recognizer that measures the cosine distance between a stored template of a gesture password and an input, and uses the reciprocal of that as a *similarity score*. Discussion of the design and implementation is beyond the scope of this paper, and has been explained in exhaustive detail in previous work [85, 118].

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 [118]. 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.

4.2.4 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 4.2a. In the first week, participants performed



(a) The 2-week process of our study.

(b) A sample schedule of creation + immediate login session for account set two.

Figure 4.2: 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.

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 [50]. 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 4.2b shows a typical schedule of a day in our study.

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

4.2.5 Password Analysis

We performed password cracking attacks to analyze text password security. We used the popular GPU-based password cracker oclHashcat 1.36 [67] 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: "basic64" and "generated2". The third one used the rule set designed by KoreLogic [71]. 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 [82]. It has been found to be effective for password cracking [123]. All attacks used the same input dictionary, a shuffled combination of different wordlists that included Google 1-gram English dataset [57], UNIX dictionary [83], RockYou leaked password dataset, and phpbb leaked password dataset. It contained 38M unique words. We followed the cracking techniques from recent literature [130] so that the results are interpretable.

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.

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.

4.2.6 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 [139], 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 [48]. For easier reading, we used tables to report the tests for multiple comparisons.

4.3 Results

Table 4.1 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 responds to it. 75% of our participants reacted to a task within 10.45 minutes.

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 4.1: 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. The completion rate was high compared with previous studies.

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

4.3.1 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 4.3 shows. The creation duration was calculated as the average time needed to create a password for one account. In two-account settings, 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.

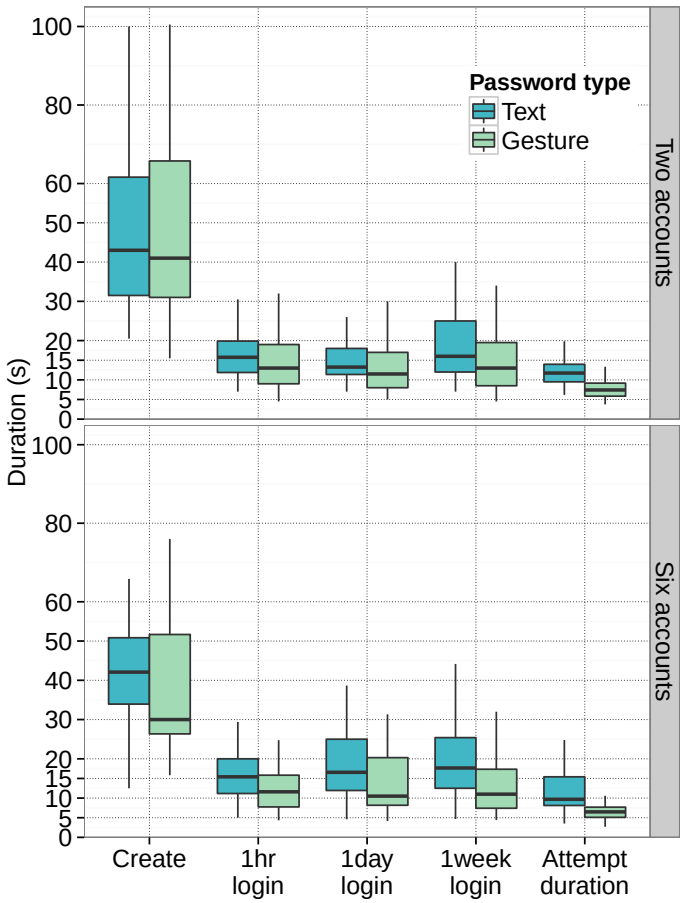


Figure 4.3: Duration data, including creation duration, log-in duration and attempt duration. Outliers are removed for clear visualization. The log-in duration is the average time needed to log in to one account successfully. The attempt duration is the time needed to perform one log-in attempt. The 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 figure shows the gesture group has an overall shorter duration than the text group in most categories.

	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 4.2: Text passwords' number of characters (left), and cracking attack result (right).The cracked result is similar to that of the weakest category of passwords being cracked under the similar experiment setup.

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 4.2 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 [130].

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 4.4. We categorized our gesture dataset by both (i) all passwords, and (ii) unique passwords.

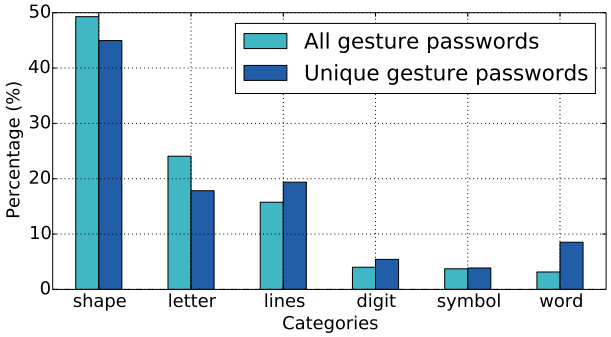


Figure 4.4: The categories of all gesture passwords created by the participants. The analysis shows the count of both all gestures passwords and only unique gesture passwords. Based on the figure, "Shapes" and "Letters" were favored by participants.

(a) (b) (c) (d) (e)

First Second Third Fourth Fifth

Figure 4.5: 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. Within "Shapes" category, participants preferred common shapes.

Overall, the most common category was the "Shapes" (49.28%), following by "Letters" (24.07%) and "Lines" (15.76%).

Figure 4.5 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 4.5e), 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 4.6.

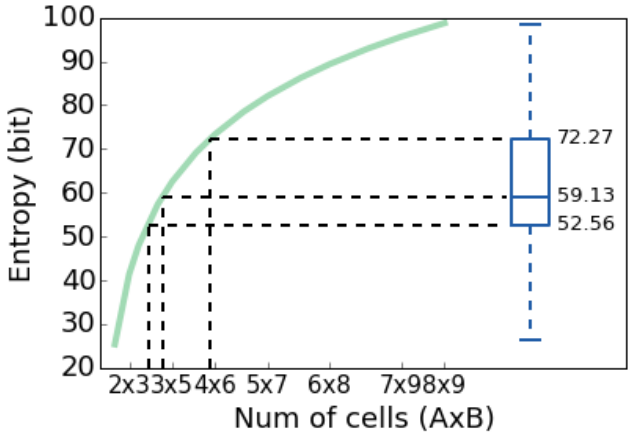


Figure 4.6: The curve displayed is the entropy of gesture passwords varied by the number of cells. Number of cells of a screen is defined as the number of cells horizontally (A) times number of cells vertically (B). The boxplot represents the entropy of our text passwords. The edge of the box represents 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. The figure shows that modeling the touchscreen to be 3x5 grid of cells matched the median entropy of our text password.

The figure also includes a boxplot of entropy of our text passwords for comparison.

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.

4.3.2 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 4.3 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

	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 4.3: 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. The results show that success rate of two groups was mostly similar across conditions.

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 4.3 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 4.4 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.

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 4.4). 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.

		Median (s)		Mean (s)		Wilcoxon rank sum test			
Account set	Log in after	Text	Gesture	Text	Gesture	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 4.4: 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.

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 4.5 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 4.3 shows the attempt duration of each log-in task.

To compare the attempt duration, we did a Wilcoxon rank sum test, and Table 4.6 shows the result of each pair-wise comparison. In all log-in tasks, the gesture group

		Median		Mean		Wilcoxon rank sum test			
Account set	Log in after	Text	Gesture	Text	Gesture	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 4.5: 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.

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 4.7. "Wrong account" referred to the 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

		Median (s)		Mean (s)		Wilcoxon rank sum test			
Account set	Log in after	Text	Gesture	Text	Gesture	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 4.6: 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.

text passwords only, meaning that the input was correct except for 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.

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 4.7. 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.

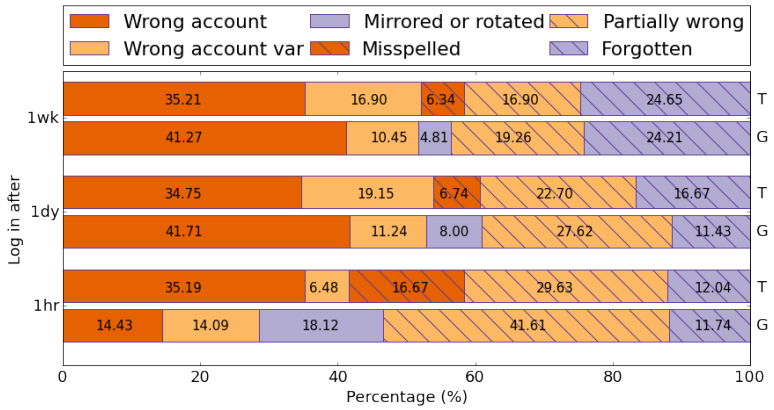


Figure 4.7: Categories of reason to fail at log-in tasks. Log in after one hour (1hr), one day (1dy) and one week (1wk) corresponds to immediate, short-term and long-term log-in tasks, respectively. Bars with label "T" are for text group and ones with "G" are for gesture group. In the legend, "wrong account var" stands for wrong account variant. The most notable thing is that the gesture group made fewer errors of "Wrong account" (and variant) than the text group in log-in tasks after one hour.

4.3.3 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 [50]. 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 4.8 showing the result. A Wilcoxon rank sum test was applied on the TLX scores, and the test result showed no statistically significant difference in TLX scores between the two groups.

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

Table 4.7: 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. In general gesture group spent less time while were willing to retry more in given-up log-in tasks.

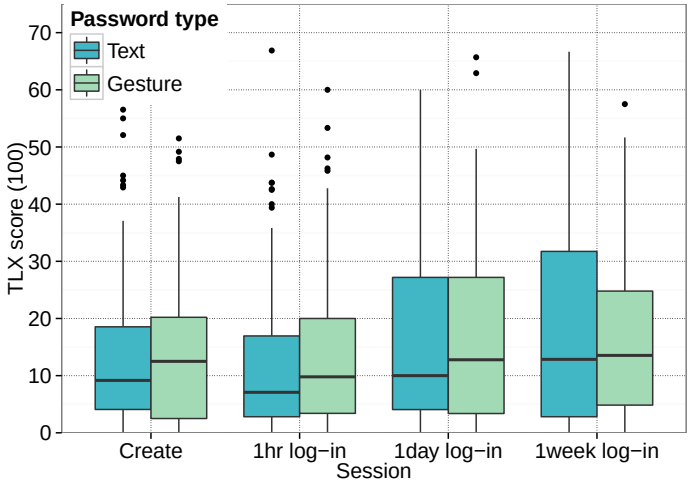


Figure 4.8: The TLX scores of two groups were similar across six individual factors. We calculated the average of the score of those factors. The ratings were similar between groups.

4.4 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 ESM for a password field study.

4.4.1 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 [62, 101, 128, 134]. 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 the similarity score between gestures from the 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 that getting used to gesture passwords was 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.

4.4.2 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 at least 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 4.7). 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 4.7). 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.

4.4.3 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 [104], 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" [118], 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 [118]. 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 [2, 47, 122]. 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 4.2).

Our proposed security metric also compared the entropy of our text and gesture passwords. For the grid of cells the screen is split into, 4x6 is still considered "low resolution", but our analysis (Figure 4.6) 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.

4.4.4 Completion Rate

The completion rate of our study was similar to that of a previous study on multi-password interference [46]. 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 [46, 134]. 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 [11]. Therefore, the result obtained should be interpreted only for relatively comparing security of the two, not measuring the absolute security of passwords.

4.5 Summary

In this chapter, we presented the first field study of free-form gesture passwords as mobile authentication method, with text passwords as the baseline. We reported a 91-participant field study, with 347 text passwords, 345 gesture passwords, and 2002 completed log-in tasks generated.

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.

This chapter also presented the analysis of the first gesture password dataset from the field. We found user-chosen gesture passwords were varied, with preferred categories being shapes (49.28%) and letters (24.07%).

Our findings contextualized the existing research on gesture passwords as well as challenge previous findings from lab studies.

Chapter 5

FACTORS IN PASSWORD MEMORABILITY

5.1 Overview of Chapter

In this chapter, we present the first systematic experiment design to study the memorability of passwords. Our design is centered on two variables: log-in frequency and password condition.

We first describe the experiment design and our procedure. We used an iterative design method. We ran a pilot study with the first version of our design. Based on the feedback and data collected, we modified the study design, and ran the formal study with the updated design. We include a subsection in the method section to explain what was revised based on the pilot study.

In the results section, we first examine data collected from the pilot study, to support the decisions to revise the study design for the formal study. Then, with the data we collected in the formal study, we examine the effect of our two variables on memorability. To measure the memorability, we look at both log-in success rate and log-in duration. We also examine whether two classic memory effects (retention effect and practice effect) exist in the memorization process of passwords.

In addition, we explore specific factors in password context to see their effect on memorability, including password reuse and password security.

5.2 Method

5.2.1 Experiment Design

Our primary objective was to study the process of people memorizing passwords. The experiment followed a common password study paradigm [26]: participants generate passwords for several accounts under different conditions, and were asked to recall multiple times at different points of time afterwards. Participants performed tasks online using a web application, instead of coming to lab. Such a design allowed participants to perform tasks at any time and anywhere, which fits the real usage of passwords better, and results in a better ecological validity.

The experiment is a 4 x 2 within-subject design with two variables: log-in frequency and password condition. Each combination of the level of the two variables would be represented as a virtual account. All participants follow the same experiment structure to manage these eight accounts. Within-subject design brought several advantages: (1) matched the password usage in realistic settings (multi-account [50] and varied log-in frequency for different accounts [68]); (2) ensure statistical power given a relatively small sample size. Note that we did not inform participants about the password conditions at all.

Controlled Variable

The log-in frequency variable is a within-subject variable indicating how frequently a participant needed to log in to a particular account. Previous study showed that people accessed their passwords in various frequencies [68], and log-in frequency played an important role in password memorability [72]. The variable has four levels: once a day, once three days, once five days and once a week. Each account would be assigned with one log-in frequency. Previous studies utilized different log-in frequency from once per hour to once per two weeks [14, 25, 27, 43]. We chose a similar frequency within the range of the previous studies.

The password condition variable controlled the conditions how the passwords were created. It has already been shown that people purposefully generate passwords with

different levels of security and behave differently for different accounts [5, 61, 94]. The purpose of this variable was to study whether such a difference exists in the memorability of passwords as well.

The password condition variable has two levels: simple and strong. The levels were differentiated with each other by two major factors: category of the account the password belonged to and security requirement the account had.

We used the account categories proposed in previous literature: identity and financial sites for strong, and content sites for simple condition [12, 61]. This categorization provided a reasonable separation of different accounts, and has been shown to match the subjective perception of importance people have regarding their accounts [61]. We have eight accounts in total, and four in each category. We designed a unique name, description, color, and interface for each account. The list of our accounts and their detailed description can be found in the appendix.

The security requirement of each account included: password meter score, naive check, and similarity to passwords of other accounts. The conditions had different requirements on each element of these security requirements. A detailed description of these requirements are shown in the next subsection.

The primary objective of including password condition as a variable is to study the effect of accounts on memorability. To the best of our knowledge, there has not existed any systematic study of such an effect. Therefore, in the setup of condition levels, we intend to make the contrast between the "simple" and "strong" level as large as possible, so that it would be easier for us to detect an effect if any.

Instead of treating account type and password security as two separate variables, we combined them into one as password condition. The variable of account types studies the effect of importance of each password has on memorability, and variable of password security is for the effect of password security. Our rationale of combining them is that password security is coupled with account types at a certain level [5,61]. We also benefit from the combination to maintain a reasonable amount of levels in our experiments and the simplicity of the design. One interesting future work could be examine passwords created for same account type under different security requirements (e.g. bank account).

Password Strength Meter

Password strength meters are well-studied, and have been revealed to have observable impact on password security as well as user behavior [45, 81, 115, 129]. Also, they have been widely deployed in use [33]. Therefore, participants are familiar with them and their behaviors are controllable. Commonly, a password-strength meter gives real-time feedback such as "Weak", "Good", "Strong" to users to indicate the security level of the password. The meters generate feedbacks based on a chosen algorithm which computes a security score. Every time they receive the update password, they display the feedback given the security score the algorithm generates for the password.

We used the password meter mechanism to enforce three security requirements for our password conditions. The first requirement was a strength score computed by the *zxcvbn* password strength estimator built by Dropbox [41]. Passwords of each condition need to at least meet the required by corresponding condition. The reason to use this particular estimator is two-fold. It is open-sourced, and has been deployed in many practical applications including Dropbox itself [40, 56]. Also, the previous study showed that compared with other implementations that primarily focus on character sets and length requirement, the zxcvbn meter measures the security based on the structure of passwords, and found to be consistent against most publicly-available password datasets [33]. The estimator computes a score based on their estimated entropy of a password. The score ranges from 0 to 4.

Our password strength meter also performed two security checks on each generated password. The naive check detected if a password of an account was the same as the corresponding account name, username and passwords of other accounts. The purpose of the naive check was to first eliminate identical passwords across accounts. Multi-account interference already has been shown to have an effect on memorability, therefore, to avoid its bias on our results, we need to ensure every participant create eight different passwords. Also, using username or account name would render the log-in tasks meaningless, which the naive check could prevent from occurring.

We did another check only for strong account category: computed edit proportion of

Level	Account category	Required meter score	Naive check	Similarity check
Simple	Identity and financial	4	Yes	No
Strong	Content	0	Yes	Yes

Table 5.1: Password strength meter setup for the two password conditions. The naive check disallowed the exact same password across multiple accounts, and the similarity check set a threshold for password reuse based on edit distance.

the password and each password of his/her other created accounts. The edit proportion was calculated as follows: 100 multiplied with edit distance, and then divided by the length of the base password. The value ranges from 0 to 100, and indicates the size of the portion a password needs to be edited in order to be the same as the other password it is compared against. The smaller the portion is, the more similar the password and the other one are. We disallowed passwords that had a score smaller than 25. We performed the same check with account name and username as well for similar reasons. To compare, half of major sites inspected by another work deployed similar checks [21].

We chose the value to discourage participants creating very similar passwords across accounts, but still allowed them to partially reuse their passwords, as password reuse is a common password management mechanism in daily life [55, 72]. We believe it plays a crucial role in the way people memorize their passwords.

Table 5.1 shows the password strength meter setup for each condition.

We realize password meters carry their own caveats. For one, they only control the lower bound for password security, and segment security into only a few levels. While it is desirable to control password security in a more granular fashion, it is difficult to achieve that level of control without harming the external validity of the experiment. We believe a password-strength meter based on a reasonable scoring algorithm strikes a balance between our research purpose and the ecological validity of the study. To compensate, we included more detailed analysis on password security after we collected the data.

We noticed there exists many other security metrics and implementations that could

be used as the scoring algorithm. Alternatives include password guidance and persuasive mechanisms [80, 115]. While they provide control over password security as well, they are rarely adopted in practical use. We argue that because we focus on understanding the memorability of passwords, it is very important for our experiment design to match the users' expected scenario of using passwords, as their behavior, and consequently memorability could be easily affected by extra mechanisms which do not exist in practical password usage scenarios.

A series of statistical metrics has been introduced recently [11, 22]. The major issue with them is that they rely on a large password dataset, making them not suitable for our case. Another popular category is crack-ability. They often use the number of guesses a particular offline password cracking algorithm needs to crack it to estimate how secure a password is [78]. While they demonstrated important aspects of the specific attack model – un-throttled offline attack model, we need a tool that could measure password security in a broader sense.

We need to point out several limitations for choosing the zxcvbn password strength estimator. The estimator itself has blind spots in terms of measuring password security: it has been found to fail to capture some insecure patterns in passwords such as word reversing (e.g. ehcsroP for Porsche) and some keyboard patterns like 1a2s3d4f5g [33].

Tasks

Our tasks included creation and log-in tasks. Participants generated a password for one account per day, regardless of which log-in frequency each account was assigned to. For each password generated, the corresponding log-in tasks were scheduled based on the log-in frequency. We designed it so that both creation and log-in tasks were differed according to their account by the interface, color and layout. The screenshot of our tasks can be found in the Appendix.

In our study, a log-in task required participants to log in their accounts with the passwords they created. Our log-in tasks did not have any limit on number of attempts participants could try, but it provided a "Give up" option for them after five consecutive failed attempts. After participants gave up a log-in task, they went through a password

recovery process in which we would show them the correct password. We realized the security issue of storing and displaying plain password text on our server; however, if we allowed participants to reset their passwords during the study, the we can not treat the memory process of new passwords generated as the same as the old ones, and therefore adding complication to the analysis. We added designs to ensure the security of the recovery process, which we will describe in the apparatus section.

Schedule

Because each participant managed multiple accounts, they performed tasks in a certain order. Letting participants create passwords sequentially provides better ecological validity than creating multiple passwords within a short amount of time. One potential drawback is it would take a much longer time for participants to complete the study in this way.

Because we have eight different accounts, there could be many different ways of ordering the incoming accounts. According to *recency effect* and *primacy effect*, if we keep a fixed order for all participants, they would likely performed better on the first and last few accounts. To avoid such bias, we generated all possible orders to form a candidate pool. Whenever a participant started the study, we picked one order randomly from the pool for this participant, and removed it from the pool.

Another common way to minimize the ordering effect is to apply latin square shuffling to the study. Latin square shuffling intends to distribute participants equally into different orderings. In our case, our sample size is much smaller than the possible ways of ordering, therefore latin square shuffling would not work.

Proactive Interference

Proactive interference is the effect of items learned previously on the current items [76]. Because we were studying passwords, the real passwords and accounts participants already have might affect the study result, which is potentially proactive interference. Therefore, it is useful to know (1) how many accounts participants already manage; (2) how many passwords they already have.

To collect such data for our study, we included relevant questions in the survey we sent to participants at the end of the study. We asked them to count the number of accounts and number of passwords they had in real life. To help them count, we displayed a list of account types, so that they could count per type.

5.2.2 Apparatus

We designed and built a web application for participants to perform required tasks. The application was written in Javascript, using Meteor framework. The application had an administration page where it showed all participants info and their tasks progress. It also provided options for researchers to send tasks given schedules of the participants. The application generated different emails depends on the type of the task. In addition, for each task, the application generated and sent a reminder email automatically if the participant had not finished it after three hours. Screenshots of the application are in the appendix.

Each task generated has an unique id. Each link participants used to access their tasks was based on its unique id. By making the link unique, and attaching a status flag to it, we controlled when participants could access each task. Each link expired 24 hour after it was sent to participants. The email recovery link also followed the same fashion. Each recovery email participants received contained a unique link. By clicking it, participants were able to see their correct password for the corresponding account. The recovery link expired after one hour. In such way, we ensured only participants themselves could proceed in their recovery process, unless they shared their emails with others. Also, expiring in one hour prevented participants from relying on the recovery process for log-in tasks.

Although we explicitly informed participants not to use any auto-save or auto-fill features to store their passwords, we added features to the application to detect and disable auto-fill password function of web browsers and password managers. The first feature was to turn the password input field to read-only, as some web browsers would only auto-fill the field if the fields are write-able. In addition, each time a page was loaded, the application checked if the password field was already filled with texts. If so,

it is likely participants were using the auto-fill function, and we alerted the participants and recorded the occurrence. Our application was not able to disable and detect all possible auto-fill functions, but we intend to minimize such behavior.

5.2.3 Procedure

At first, participants were introduced to the study and asked for consent to participate in the experiment. After consenting, we asked participants to watch the introduction video, and ask questions if any. The video explained the study and how to perform tasks using the web application.

For the next month, participants would receive emails to notify them new tasks. Each email contained a link for them to access the web application. Links expired after 24 hours, so that to ensure they respond in time. Each link was unique and assigned to one participant, and participants could only access tasks through valid links. Participants performed tasks during their daily lives. The task was either registration or log-in.

A registration task asked participants to create username and password. We required them (1) not to use their real passwords; (2) create a secure password; (3) not to write down or store passwords. We have no specific requirement on passwords created, although password condition variable was applied to ensure the level of security as mentioned previously.

A log-in task required participants to log in the account with the password they created. The study did not limit the time or the amount of attempts one could spent on each log-in task. As mentioned in an earlier section, we provided a "Give up" option for participants after five consecutive failed log-in attempts, which would offer them a chance to see their correct passwords and memorize them again.

Upon the end of the study, we sent out an online survey to collect feedback, and they came in again for a brief interview for the experience of the tasks, and got compensated.

Revisions Based On Pilot Study

As mentioned previously, we followed an iterative methodology to improve our experiment design. We ran a pilot study to collect feedback and revised our initial study design. After the pilot study, we ran our formal study in two batches. The first batch used the revised design after the pilot study, with the aim of exploring if our designed variables were properly explained the memory effect, or do we need to include more possible variables. Based on the result of first batch study, we would again revise the study design, and derive the model eventually. Here we describe the revisions we made for the formal study, and justify such revisions in the discussion section.

The major difference between the pilot and the formal study was that, in pilot study, we did not add the requirement of naive check and similarity check in the password condition variable. We added it to the formal study because of what we found from the pilot study. In addition to the password check, we made a few other minor modifications. In pilot study, emails of different accounts and tasks had same title. We modified our system such that each account had a set of unique email templates, and each email contained a unique timestamp to distinguish from each other.

5.2.4 Analysis

To examine password reuse, we used edit proportion, which is a normalized version of edit distance [84]. For two passwords, we calculated first the edit distance, and then normalized it by dividing it using the length of the longer one. The edit proportion ranges from 0 to 100. The larger the value, the less similar the two passwords are.

Because participants created the set of accounts in an order, and each account has a different frequency, accounts of each participant had different amounts of log-in tasks. Accounts that were created earilier or with a higher log-in frequency resulted in more log-in tasks than others. The difference in the amount of tasks could possibly bias our results. Therefore, for some studies, we need to select only the first few tasks of each account within each participant. We created a dataset called *equal-task-amount* dataset, where we only selected the first two log-in tasks of each account. We describe

the detail of this dataset for our studies in the result section, and mention it specifically if the analysis is done with it instead of the entire dataset.

In addition to the security estimation of the password meter we deployed, we performed several cracking attacks to measure the password security. Similar to the previous chapter, we used the GPU-based password cracker oclHashcat [67] to generate rule-based attacks, based on four popular sets of rules: "T0XlCv1", "rockyou-30000", "generated2" and "dive". These rule sets were generated by community experts based on previously-leaked password datasets. For input dictionary, we added crackstation human-only passwords [30] to the wordlists used in the previous chapter, resulting in an input dictionary of 150M unique words. To compare the cracking performance relatively, we repeated the same four attacks on several datasets: two password datasets from this study (one from pilot, one from formal study), one dataset from chapter three and one dataset from chapter four. Additionaly, we repeated one more time on the fromal study dataset with the variations of account names of our study (37 unique entries in total) being included into the input dictionary. Because our study introduced eight virtual accounts with distinct names and designs, we would like to see the resistance of our dataset against such targeted attacks.

5.2.5 Participants

For the pilot study, we recruited seven participants from our department, or people we have known. Four of them are male and three are female, with an average age of 28 years old ($Std = 4.31, Mdn = 27$). Three hold a Bachelor degree, while the other four hold a graduate degree.

For the formal study, we recruited participants by posting fliers, sending emails to campus mailing lists, and posting in online forums. Participants were required to be at least 18 years old, and were familiar with web services.

We have recruited 10 participants for the first batch of formal study. 55.56% are female and 44.44% are male. The average age of participants is 25 years-old ($Std = 3.28, Mdn = 23$). 44.44% of them have a Bachelor degree, 44.44% have a graduate degree and 11.12% have an Associate degree.

5.3 Results

5.3.1 Pilot Study

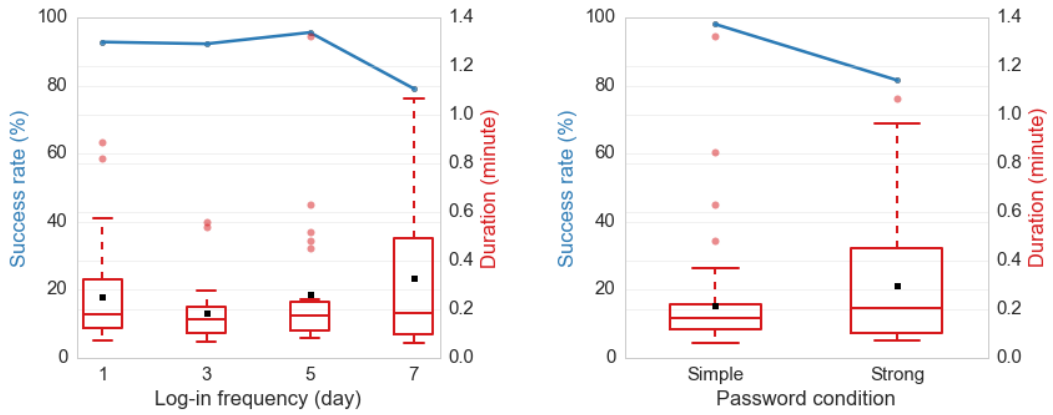
Our seven pilot participants created 56 accounts, and performed 486 log-in tasks. The overall completion rate is 86.31%, with 86 tasks expired. The average response time to a task is 4 hours 36 minutes, with a minimum of 21 seconds and a maximum of 22 hours 54 minutes.

The duration needed to create passwords under two password conditions differed for our pilot. The overall average creation duration was 49.94 seconds ($Std = 43.05, Mdn = 35.44$). That of simple and strong condition were 40.64 ($Std = 26.87, Mdn = 32.63$) and 58.91 seconds ($Std = 53.31, Mdn = 42.85$), respectively.

We found participants reused passwords, and that affected memorability. As mentioned in an early section, we did not check for the password reuse case in creation tasks. Therefore, among the 56 passwords, there were 47 unique passwords. There were two participants who reused their passwords entirely: one used the same password for all accounts, the other one used six passwords for eight accounts. Even for participants that created eight distinct passwords, some of them reused part of the passwords: the average edit proportion per participant was only 59.31%. As described in method section, edit proportion of a pair of passwords was the proportion of a password needed to be changed to match the other password. Moreover, we found accounts with an edit proportion smaller than 50% have a higher log-in success rate on average (96%) than those with a larger proportion (81%).

We found that both frequency and password condition have effects on the log-in success rate. We used the *equal-task-amount* dataset for this analysis. It contains 51 accounts of seven participants, and 102 log-in tasks. The log-in success rate of simple and strong condition were 98.07% and 81.63%, respectively. Figure 5.1 shows the detailed success rate and duration per frequency and condition. The trend in both figures are observable: success rate decreases when frequency becomes lower, while it decreases when password condition is strong, compared with simple condition.

The data from the pilot study did not reveal an observable retention effect or practice



(a) (b)

Figure 5.1: The plot of log-in success rate of different log-in frequency (a) and password condition (b). Both two factors show a visible effect on log-in success rate. In both figures blue lines indicate the success rate, with the Y-axis being on the left; the red boxplots indicate log-in duration data, with the Y-axis being on the right. For duration boxplots, the black squares indicate the mean value. According to the figure, the log-in success rate decreased when the log-in frequency became lower, or the password condition changed from simple to strong.

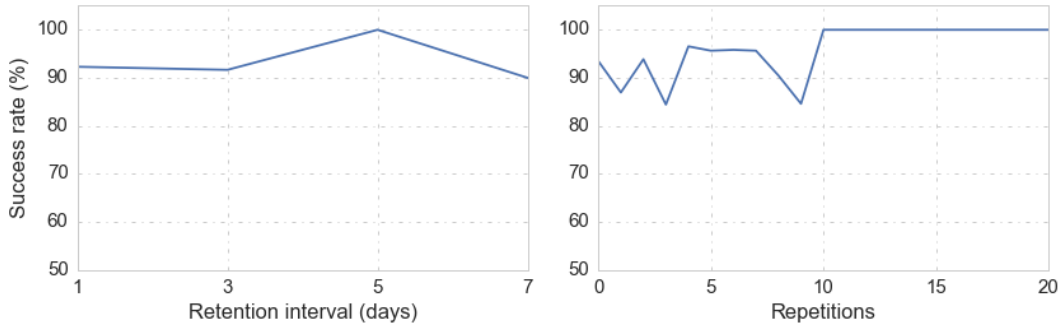


Figure 5.2: Log-in success rate of the first log-in task for all accounts grouped by frequency (retention effect, left figure), and log-in success rate of all the log-in tasks in order as repetitions (practice effect, right figure). Neither curves show an obvious pattern, indicating no retention effect or practice effect in our data.

effect of password memorability. Figure 5.2 shows both curves, with the metric being log-in success rate. Both curves were oscillating without any obvious pattern. The retention effect did not seem to apply on the password memorability, according to this result. The practice effect showed that the success rate converged to 100% after 10 consecutive repetitions.

Seven participants responded our mid-point feedback survey, and their subjective rating of task difficulty was 4.57 out of 10 in average. To compare, in exit survey, they rated 3.57 for the same question.

Six of them answered "Yes" to the question "Do you find it easy and fast to identify our emails and access our tasks". The remaining one explained the reason he chose "No": "the tasks are bundled together in one thread...after a while I lose track of which task I completed, since the links look the same to me". Relevant to easier recognition of the email, another participant suggested that "the appearance of the email should be tailored for each task".

Six indicated they were aware of the feature that every task expires 24 hours after arrival. One participant explained why he missed tasks: "read the email when I don't have time to do the study, then later on forget to check the read email".

According to the exit survey, one participant admitted to using password manager,

another to using web browser to save passwords, and no participants wrote passwords down. Only 42.9% (3) of them agreed that "passwords with stronger requirements harder to memorize", and 71.4% (5) agreed that "more frequently used passwords were easier to memorize".

We also asked questions regarding participants' password usage in their daily life, outside of the study. 71.4% (5) of them admitted they reuse passwords in daily life. According to the exit survey, 42.9% (3) chose "once per day" as the frequency of their most-frequently-used account, and 57.1% (4) chose "less than once per month" as that of their least-frequently-used account.

5.3.2 Formal Study

In the formal study, 10 participants created 80 accounts and performed 614 log-in tasks in total. The overall completion rate was 96.39%. Each account in our study generated 8.0 tasks in average ($Std = 7.17, Mdn = 4$), and each participant completed 72.0 tasks on average ($Std = 8.54, Mdn = 74$). Tasks sent in the formal study have an average response time of three hours 52 minutes, with a range from 22.8 seconds to 23 hours 57 minutes. 35.59% of tasks triggered the reminder email function as they have not been completed three hours after arrival.

Creation

The average duration participants needed to generate a password under simple and strong condition were 71.95 ($Std = 76.53, Mdn = 47.09$) and 108.81 seconds ($Std = 98.20, Mdn = 82.32$), respectively. Figure 5.3 shows a distribution of creation duration for both conditions.

Our participants created 80 unique passwords. Table 5.2 shows a detailed description of our passwords. Figure 5.4 shows the boxplot of character counts per password for two password conditions. The results indicates that our passwords are non-trivial passwords with complex structures.

We found the security measure of our passwords by the password meter were not very different between the two conditions. The average score for strong condition is

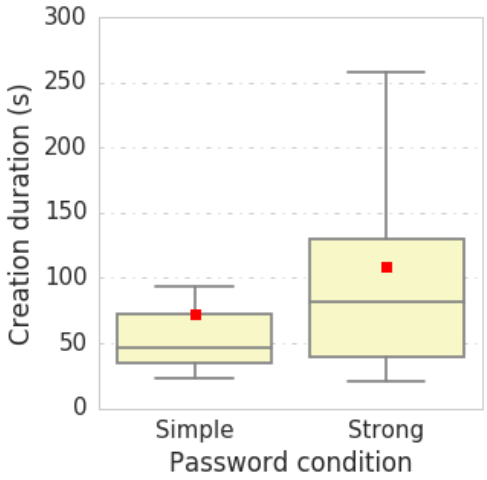


Figure 5.3: The boxplot of creation duration for two password conditions. Red squares indicate the mean value. It shows that participants spent more time in generating passwords under strong condition than the simple condition.

	Length	Lowercases	Uppercases	Digits	Symbols
Mean	14.05	9.83	0.93	2.84	0.45
Std	3.68	4.37	1.51	2.79	1.25
Mdn	13.5	10.5	0	3	0

Table 5.2: General password statistics, including password length, amount of lowercase letters, uppercase letters, digits and symbols per password. The results show that passwords generated in our study are non-trivial ones with complex structure.

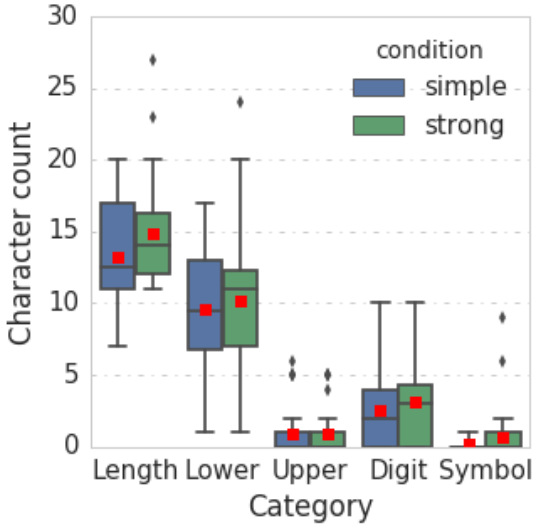


Figure 5.4: The boxplot of password length and characters count for two password conditions. The characteristics of passwords in two conditions are not very different from each other.

4.0, and that for simple condition is 3.25 ($Std = 0.98, Mdn = 4$). More than half of passwords in simple condition (55.0%) have the same meter score as that of strong condition. In addition, the average entropy of the simple and strong condition were 10.28 bits ($Std = 3.51, Mdn = 10.44$) and 12.07 bits ($Std = 1.96, Mdn = 11.79$).

We found that passwords generated using this experiment design were much more resistant against password cracking attacks, compared with that generated from previous studies. Figure 5.5 shows the result of four different cracking attacks on four password datasets: pilot study, formal study, text entry and field study. The first two were from memorability experiments, and the other two were from chapter three and four, respectively. A detailed description of the cracking attacks setup and datasets can be found in the method section. One can see that, cracked percentage of datasets from the memorability design were below 30%, while that of others were above 50% with the highest reaching beyond 70%.

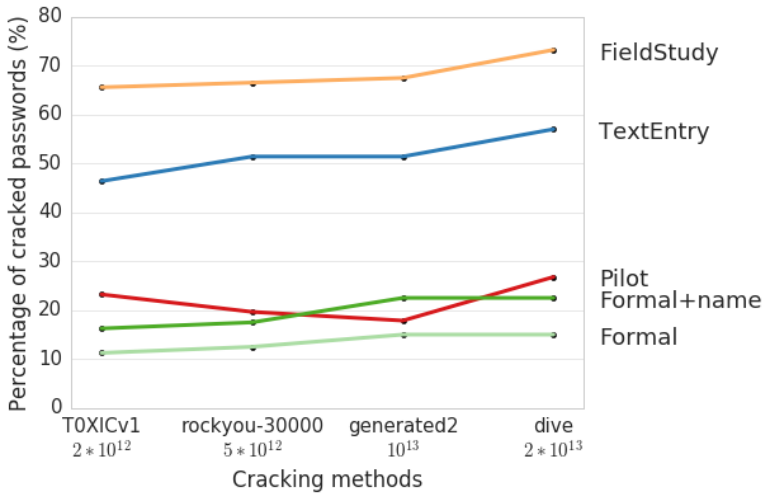


Figure 5.5: The result of our cracking attacks. X-axis represents different cracking methods and their total guesses, and each line represents a password dataset. "Field study" is from chapter four, "TextEntry" is from chapter three. "Pilot" and "Formal" are from this chapter. "Formal+name" used the same dataset as "Formal", but adding variations of account names in our study to the input dictionary to create a targeted attack, as mentioned in the method section. The figure shows that datasets from this study have better resistance against cracking than others, and our targeted attack cracked nearly 10% more of the "Formal" dataset than the regular attack.

The formal study did not have naive password reuse case as mentioned in the method section. The average edit distance between passwords for each participant was 9.58 ($Std = 2.49, Mdn = 9.32$), and edit proportion was 73.03% ($Std = 22.95\%, Mdn = 79.33\%$). To provide context, the average length per password was 14.05 in our study.

Log-in

Our 10 participants gave up 49 tasks, with an overall log-in success rate of 92.02%. The average duration one spent on single log-in task was 28.36s ($Std = 53.65, Mdn = 15.3$), 23.67s for successful log-in tasks and 97.17s for tasks that they eventually gave up. The *equal-task-amount* dataset for the formal study contained 78 accounts of 10 participants, and 156 log-in tasks. The log-in success rate for this dataset was 80.12%, and log-in duration was 46.59s on average ($Std = 91.87, Mdn = 23.44$).

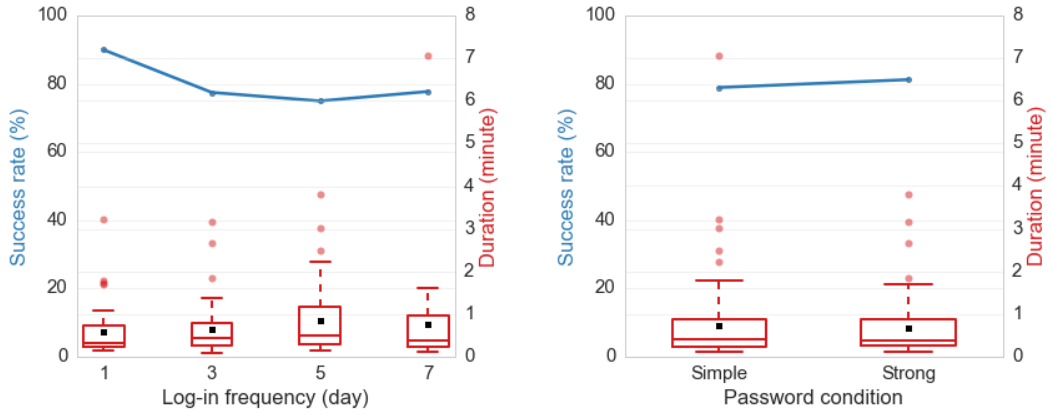
Figure 5.6 shows the detailed success rate and duration for each log-in frequency and password condition. The figure was computed using *equal-task-amount* dataset. From the figure, we can see a trend of decreasing success rate as the log-in frequency decreases, but the password condition seems not have a clear effect on the change of rate. In addition, there did not seem to exist any observable effect of the two variables on log-in duration.

We found a possible practice effect from our data, although the retention effect was less obvious. Figure 5.7 shows both curves, with the log-in success rate as the metric. The practice curve converges to 100% after 18 repetitions. However, retention curve remains nearly flat, and does not seems to be affected by the log-in frequency.

Exit Survey

In their exit survey, participants rated the tasks with an average difficulty of 4.44/10 ($Std = 1.59, Mdn = 5$). None of them reported to write down passwords, or use password manager or browswer auto-save functions. We asked them to estimate their own log-in success rate, and the average estimation was 83.33%.

When asked about strategies or patterns they used to memorize passwords in our



(a) (b)

Figure 5.6: Figure (a) is the log-in success rate and duration of different log-in frequency; Figure (b) is that for different password conditions. In both figures blue lines indicate success rate, with Y-axis being on the left; red boxplots indicate log-in duration data, with Y-axis being on the right. For duration boxplots, the black squares indicate the mean value. The figure shows a decreasing trend in success rate when log-in frequency lowers, but no visible effect of password condition on the rate.

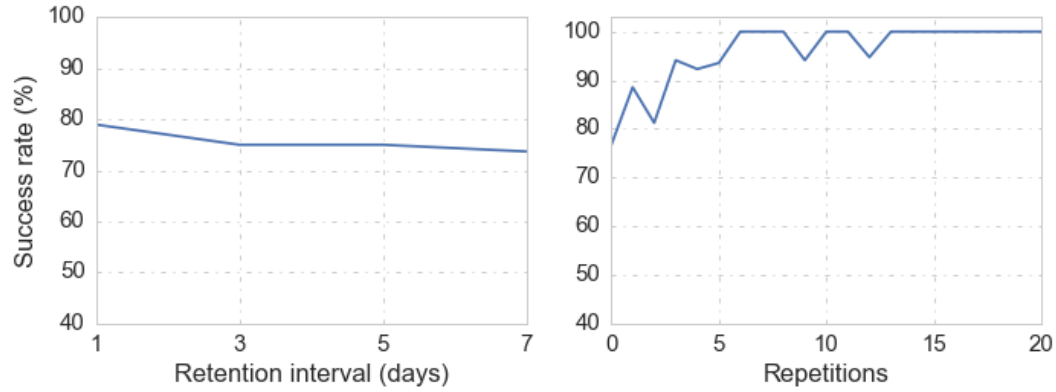


Figure 5.7: Log-in success rate of the first log-in task for all accounts grouped by frequency (retention effect, left figure), and log-in success rate of all the log-in tasks in order as repetitions (practice effect, right figure). The retention effect in the left figure is very subtle, and we can see a visible practice effect in the right figure.

study, 44.44% mentioned their passwords included the name of accounts. Two participants reported one of their passwords in the study was similar or same to the ones they have in the daily life, outside of the study. Another participant reported two in the same question. Others did not have such cases.

Only 33.33% of participants believed different strengths of passwords would affect their memorability, and 66.67% agreed different log-in frequency would have effect on memorability.

Regarding their daily-life password usage, participants reported they manage 13.3 accounts and 7.4 unique passwords on average. 88.89% of them admitted they reused passwords in daily life. 55.56% of participants reported "Multipe times per day" as the most-frequent access of their accounts, and 55.56% of them reported "Less than once per month" as the least-frequent access.

Factors To Affect Memorability

In this section we examine the relationship of different factors with password memorability, which is represented as the log-in success rate or binary log-in status (succeed or failed). We computed the result using the *equal-task-amount* dataset.

We found several per-account factors that affected memorability. We computed the Pearson correlation coefficient r for each factor we found in the study with password memorability. Table 5.3 shows the detailed results. Factors of log-in tasks (attempts, duration, task completed) showed strong effects, as well as the experiment variable log-in frequency. However, we did not observe the effect of password condition variable.

We also found a few per-task factors that have an effect on memorability. Table 5.4 shows the detailed correlation of each per-task factor and log-in task status (succeed or not). Attempts and duration of each task are shown as strong factors, as well as log-in frequency. The sequence number of the task also affected the outcome: the later the task is, the greater the chance of a successful log-in could be.

Factor	r	p
Log-in attempts*	-0.43759	0.00006
Creation duration*	-0.24352	0.03168
Log-in duration*	-0.24205	0.03276
Guesses needed to crack	-0.21581	0.07065
zxcvbn entropy	0.13849	0.22658
Log-in frequency	-0.13005	0.25644
Num of uppercase letters	0.09898	0.38860
Task completed	0.09552	0.40547
Elapsed time since last log-in	-0.09505	0.40780
Num of digits	0.09296	0.41821
Is cracked	-0.07598	0.50851
Accounts	-0.07242	0.52861
Password length	0.06860	0.55063
zxcvbn crack time	0.06655	0.56267
Account order	0.06413	0.57697
zxcvbn score	0.04926	0.66843
Num of symbol	-0.04595	0.68957
Password condition	0.03357	0.77047
Edit distance	0.03103	0.78740
Num of expired tasks	0.02513	0.82716
Num of lowercase letters	-0.02015	0.86097
Edit proportion	-0.01210	0.91628

Table 5.3: Table of Pearson's correlation coefficient r and corresponding p value computed between each per-account factor and the log-in success rate. Factors such as num of symbols are the amount of characters in a single password. Factors such as log-in attempts are the average value of each account. Factors with prefix "zxcvbn" were computed by the zxcvbn password strength meter. As indicated, attempts and duration factors show strong correlation with log-in success rate.

Factor	r	p
Attempts*	0.55567	0.00000
Duration*	0.39068	0.00000
Frequency*	0.28786	0.00000
No. of the task*	-0.23331	0.00000
Time since creation*	-0.10601	0.00862
Time since last log-in task*	0.09879	0.01441
Password condition	0.01018	0.80142
Account	-0.00977	0.80929

Table 5.4: Table of Pearson's correlation coefficient r and corresponding p value computed between each per-task factor and the log-in status of the task. "No. of the task" refers to the order of this task. "Account" differentiates accounts from each other. Attempts, duration, the order of the task and log-in frequency show strong correlation with log-in status.

5.4 Discussion

5.4.1 Limitations

Our study only focused on user-chosen passwords, excluding other password types such as system-assigned passwords. The memory process of system-assigned passwords is different from user-chosen ones, because generation effect is applied when people generate their own passwords [135], while not the case if passwords are assigned to people. Consequently, new methodologies might be required. Therefore, while it would be interesting to explore the memorability of system-assigned passwords, it is beyond the scope of this work.

5.4.2 Pilot Study

We observed a very high log-in success rate from the pilot study: 94.44% overall and 98.07% for simple condition. This might due to the fact that tasks were easy in our pilot design (subjective rating of task difficulty was 4.6/10 on average), and thusly creating a ceiling effect on our variables.

It is likely that password reuse made log-in tasks easier for participants according

to our analysis. We even had pilot participants that used only one password for all eight accounts. Reusing passwords also ignored our study design, in which we aimed at exploring the multi-account interference of eight different accounts. Therefore, we first disallowed entire password reuse for all accounts. In addition, we prevent the naive partial reuse for accounts of the strong condition, by requiring the password created has a minimum edit proportion of 0.25 with other accounts.

The completion rate of pilot study was also low (86%). Based on the mid-way survey feedback from the pilot study, it is likely that participants sometimes did not finish tasks immediately after reading the email, and then just forgot about it later on. Therefore, we added a reminder function to each task in the formal study: sending a reminder email to participants for each task that was still not complete after three hours. As a result, the completion rate increased to 95% in formal study. Participants had complained in the survey that emails of different tasks were bundled together because they had the same title, making it difficult to differentiate tasks and know whether each task was completed or not. Therefore, we modified it so that emails from different accounts and tasks had unique title and content.

More importantly, we did observe that there exists such an effect of log-in frequency and password condition on log-in success rate. In particular, accounts of the strong condition have an overall higher log-in success rate than that of simple condition, and tasks with lower frequency (e.g. seven-day-login) have a lower success rate than ones with a higher frequency. However, such effects are not observable for log-in duration.

5.4.3 Formal Study

Our study results shows that the log-in frequency has a strong effect on password memorability, both in terms of a password and a single log-in task. In particular, per-account log-in success rate fell to 70% from 90% when the frequency was reduced to three days from one day (Figure 5.6a), which then remained largely flat for lower frequencies. Such results suggested that the effect might primarily exist in higher frequencies, and largely disappears once frequency reduces to several days per log-in. Consequently, it means people remember the frequently-accessed accounts the best,

but for accounts with lower log-in frequencies, different frequencies would not affect the likelihood of them forgetting the password much. It also suggests that in future studies we could further expand the range of our log-in frequency variable, to be even more frequent such as multiple times per day, or less frequent such as once per two weeks.

The other variable, password condition, did not show effect on memorability. One of the major design issues which might cause this was noticed from both the pilot and formal studies: the password strength meter did not provide sufficient control over password security. The reason is that they only allow control on the lower-bound of password security: participants were asked to generate passwords with a password meter score larger than the minimum requirement of each password condition. There is no upper bound requirement, meaning that even for the simple condition participants were free to generate complex passwords. This is understandable in real world scenarios because their purpose is usually to prevent users generating weak passwords. However, the purpose of our study was to enforce different levels of security for different password conditions so that we can compare their effect on memorability. As a result, we found passwords in different conditions were not necessarily different from each other in terms of security. Therefore, it is important to have a proper control of upper-bound of password security in our future design, while maintaining the ecological validity.

In addition, our design included eight distinct accounts, and assigned different conditions to them based on the previous findings that people categorized their accounts mentally [61]. Participants in our study were not informed that there existed differences of any kind for those accounts. Therefore, it is possible that participants did not treat the eight accounts differently, and thusly generated similar passwords for all accounts. One evidence was that our security estimates were similar between conditions and across accounts. Another possible clue was that while our requirement of meter score for simple condition was 0, half of the passwords in the simple condition were scored 4 - the highest score of the password strength meter. Therefore, including the subjective perception of each account and password as a factor in password memorability could be one of the future studies.

The behavior of password reuse also did not pose any effect on the password memorability. Neither edit distance nor proportion had notable correlation with the log-in success rate. However, in the pilot study, reuse metrics did have an effect on the log-in success rate, that is smaller edit distances and proportions associated with higher success rate. We postulate that the effect we observed from pilot study came from naive reuse cases, because we disallowed that in the formal study. Naive reuse cases included either reusing the entire password, or same password with minor modification. Such results indicated that the effect of password reuse on password memorability is limited to the cases of naive reuse. Considering password reuse is a common practice in password management, it is useful to study the naive reuse cases in mode details. For example, could there be a quantitative metric to compute the memorabilty given the level of password reuse?

5.5 Summary

This chapter presented an experimental methodology to systematically examine factors that influence password memorability. In addition, two studies were designed and executed in an iterative fashion to study password memorability as well as refine the experiment design.

We found that both log-in frequency and naive password reuse affected password memorability, when log-in success rate was used as a metric. However, the effect of password-related factors (password condition, security estimate, password characteristics such as length) were found to be very limited. Our results provided initial results for more future research to eventually quantify the memorability of passwords.

Chapter 6

CONCLUSIONS

The fundamental topic that this thesis focused on is to understand issues of usable security in mobile authentication experiences.

We first examined how the experience differed for mobile and traditional platforms. Mobile-specific issues such as switching keyboard layouts altered participants' behavior and consequently the passwords generated. Our results suggest that it is necessary to consider mobile platforms separately in terms of authentication design, instead of just migrating the legacy design from traditional computers.

As the next step, our field study explored the strength of mobile-specific alternative as an authentication scheme on mobile platforms. *Free-form gesture passwords* were considered not only natural to touchscreen interactions from the aspect of form factors, but also suitable for the fast and frequent mobile contexts. Our findings revealed that the core advantage it carries is the speed: creating passwords and authenticating faster. Saving several seconds is critical when the duration of mobile interactions are also a matter of seconds sometimes.

Finally, we looked at a rarely-studied topic of passwords: quantifying the factors of passwords memorability. In our results, frequency-related factors showed notable effect. Surprisingly, our results indicated that password-related factors have limited effect on memorability: long passwords or ones with a higher entropy estimate did not necessarily have a lower log-in success rate. Such observations might be the key to find the balance between security and memorability for passwords. As our two studies were conducted in a relatively small scale, larger studies with a longer time frame are necessary to confirm our findings.

References

[1] A. Adams, M. A. Sasse, and P. Lunt. Making passwords secure and usable. In *Proceedings of HCI on People and Computers XII*, HCI 97, pages 1–19, London, UK, UK, 1997. Springer-Verlag.

[2] F. Alt, S. Schneegass, A. S. Shirazi, M. Hassib, and A. Bulling. 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, pages 316–322, 2015.

[3] J. R. Anderson and L. J. Schooler. Reflections of the environment in memory. *Psychological Science*, 2(6):396–408, 1991.

[4] A. J. Aviv, K. Gibson, E. Mossop, M. Blaze, and J. M. Smith. Smudge attacks on smartphone touch screens. In *Proc. WOOT'10*, pages 1–7. USENIX Association, 2010.

[5] D. V. Bailey, M. Dürmuth, and C. Paar. *Security and Cryptography for Networks: 9th International Conference, SCN 2014, Amalfi, Italy, September 3-5, 2014. Proceedings*, chapter Statistics on Password Re-use and Adaptive Strength for Financial Accounts, pages 218–235. Springer International Publishing, Cham, 2014.

[6] J. M. Barnes and B. J. Underwood. "fate" of first-list associations in transfer theory. *Journal of experimental psychology*, 1959.

[7] L. F. Barrett and D. J. Barrett. An introduction to computerized experience sampling in psychology. *Social Science Computer Review*, 2001.

[8] R. Biddle, S. Chiasson, and P. Van Oorschot. Graphical passwords: Learning from the first twelve years. *ACM Comput. Surv.*, pages 19:1–19:41, 2012.

[9] J. Blocki, S. Komanduri, L. Cranor, and A. Datta. Spaced repetition and mnemonics enable recall of multiple strong passwords. *arXiv preprint arXiv:1410.1490*, 2014.

[10] J. Bonneau. The science of guessing: Analyzing an anonymized corpus of 70 million passwords. In *2012 IEEE Symposium on Security and Privacy*, pages 538–552, May 2012.

[11] J. Bonneau. 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, pages 538–552, Washington, DC, USA, 2012. IEEE Computer Society.

[12] J. Bonneau and S. Preibusch. The password thicket: Technical and market failures in human authentication on the web. In *WEIS*, 2010.

[13] J. Bonneau, S. Preibusch, and R. Anderson. A birthday present every eleven wallets? the security of customer-chosen banking PINs. In *Proc. of FC'12*, 2012.

[14] J. Bonneau and S. Schechter. Towards reliable storage of 56-bit secrets in human memory. In *Proc. USENIX Security 14*, pages 607–623, San Diego, CA, 2014. USENIX Association.

[15] J. Bonneau and E. Shutova. Linguistic properties of multi-word passphrases. In *Proceedings of the 16th International Conference on Financial Cryptography and Data Security*, FC'12, pages 1–12, Berlin, Heidelberg, 2012. Springer-Verlag.

[16] G. E. Briggs. Acquisition, extinction, and recovery functions in retroactive inhibition. *Journal of Experimental Psychology*, 1954.

[17] J. Brodkin. 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/`.

[18] S. Brostoff and M. A. Sasse. "ten strikes and you're out": Increasing the number of login attempts can improve password usability. In *Proc. of CHI 2003 Workshop on HCI and Security Systems*, 2003.

[19] W. E. Burr, D. F. Dodson, E. M. Newton, R. A. Perlner, W. T. Polk, S. Gupta, and E. A. Nabbus. Sp 800-63-1. electronic authentication guideline. Technical report, National Institute of Standards & Technology, 2011.

[20] D. Buschek, A. De Luca, and F. Alt. Improving accuracy, applicability and usability of keystroke biometrics on mobile touchscreen devices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1393–1402, New York, NY, USA, 2015. ACM.

[21] X. D. C. D. Carnavalet and M. Mannan. A large-scale evaluation of high-impact password strength meters. *ACM Trans. Inf. Syst. Secur.*, 18(1):1:1–1:32, May 2015.

[22] C. Castelluccia, M. Dürmuth, and D. Perito. Adaptive password-strength meters from markov models. In *Proc. of NDSS'12*, 2012.

[23] ChaosComputerClub. Chaos computer club breaks apple touchid, 2013. Retrieved May 3 2015 from `http://www.ccc.de/en/updates/2013/ccc-breaks-apple-touchid`.

[24] I. Cherapau, I. Muslukhov, N. Asanka, and K. Beznosov. On the impact of touch id on iphone passcodes. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 257–276, Ottawa, July 2015. USENIX Association.

[25] H.-Y. Chiang and S. Chiasson. Improving user authentication on mobile devices: A touchscreen graphical password. In *Proc. MobileHCI '13*, pages 251–260. ACM, 2013.

[26] S. Chiasson, A. Forget, R. Biddle, and P. C. van Oorschot. Influencing users towards better passwords: Persuasive cued click-points. In *Proc. BCS-HCI '08*, pages 121–130. British Computer Society, 2008.

[27] S. Chiasson, A. Forget, E. Stobert, P. C. van Oorschot, and R. Biddle. Multiple password interference in text passwords and click-based graphical passwords. In *Proc. CCS'09*, pages 500–511. ACM, 2009.

[28] S. Chowdhury, R. Poet, and L. Mackenzie. Passhint: Memorable and secure authentication. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 2917–2926, New York, NY, USA, 2014. ACM.

[29] G. D. Clark and J. Lindqvist. Engineering gesture-based authentication systems. *Pervasive Computing, IEEE*, pages 18–25, Jan 2015.

[30] CrackStation. Crackstation's password cracking dictionary, 2009. Accessed: 2016-05-01 from `https://crackstation.net/buy-crackstation-wordlist-password-cracking-dictionary.htm`.

[31] M. Csikszentmihalyi and R. Larson. Validity and reliability of the experience-sampling method. *The Journal of nervous and mental disease*, 1987.

[32] M. Csikszentmihalyi, R. Larson, and S. Prescott. The ecology of adolescent activity and experience. *Journal of Youth and Adolescence*, 1977.

[33] X. de Carné de Carnavalet and M. Mannan. From very weak to very strong: Analyzing password-strength meters. In *Network and Distributed System Security (NDSS) Symposium 2014*. Internet Society, February 2014.

[34] A. De Luca, A. Hang, F. Brudy, C. Lindner, and H. Hussmann. 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, pages 987–996, 2012.

[35] A. De Luca, A. Hang, E. von Zezschwitz, and H. Hussmann. 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, pages 1411–1414, 2015.

[36] A. De Luca, M. Harbach, E. von Zezschwitz, M.-E. Maurer, B. E. Slawik, H. Hussmann, and M. Smith. 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, pages 2937–2946, 2014.

[37] A. De Luca and J. Lindqvist. Is secure and usable smartphone authentication asking too much? *Computer*, 48(5):64–68, May 2015.

[38] A. De Luca, E. von Zezschwitz, N. D. H. Nguyen, M.-E. Maurer, E. Rubegni, M. P. Scipioni, and M. Langheinrich. Back-of-device authentication on smartphones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2389–2398, 2013.

[39] M. Dell'Amico and M. Filippone. Monte carlo strength evaluation: Fast and reliable password checking. In *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pages 158–169, New York, NY, USA, 2015. ACM.

[40] DropBox. dropbox/zxcvbn: a realistic password strength estimatior. Retrieved April 4 2016 from `https://github.com/dropbox/zxcvbn`.

[41] DropBox. zxcvbn: realistic password strength estimation. Retrieved April 4 2016 from `https://blogs.dropbox.com/tech/2012/04/zxcvbn-realistic-password-strength-estimation/`.

[42] M. Duggan and A. Smith. Cell internet use 2013, sep 2013. Pew Research Centers Internet & American Life Project.

[43] H. Ebbinghaus. *Memory: A contribution to experimental psychology*. Teachers college, Columbia university, 1913.

[44] S. Egelman, S. Jain, R. S. Portnoff, K. Liao, S. Consolvo, and D. Wagner. Are you ready to lock? In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 750–761, 2014.

[45] S. Egelman, A. Sotirakopoulos, I. Muslukhov, K. Beznosov, and C. Herley. 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, pages 2379–2388, New York, NY, USA, 2013. ACM.

[46] K. M. Everitt, T. Bragin, J. Fogarty, and T. Kohno. A comprehensive study of frequency, interference, and training of multiple graphical passwords. In *Proc. CHI '09*. ACM, 2009.

[47] S. Fahl, M. Harbach, Y. Acar, and M. Smith. On the ecological validity of a password study. In *Proc. of SOUPS'13*, 2013.

[48] A. Field, J. Miles, and Z. Field. *Discovering statistics using R*. SAGE Publications, 2012.

[49] E. A. Fleishman and J. F. Parker Jr. Factors in the retention and relearning of perceptual-motor skill. *Journal of Experimental Psychology*, 64(3):215, 1962.

[50] D. Florencio and C. Herley. A large-scale study of web password habits. In *Proc. WWW '07*, pages 657–666. ACM, 2007.

[51] D. Florêncio, C. Herley, and B. Coskun. Do strong web passwords accomplish anything? In *Proceedings of the 2Nd USENIX Workshop on Hot Topics in Security*, HOTSEC'07, pages 10:1–10:6, Berkeley, CA, USA, 2007. USENIX Association.

[52] D. Florêncio, C. Herley, and P. C. Van Oorschot. Password portfolios and the finite-effort user: Sustainably managing large numbers of accounts. In *Proceedings of the 23rd USENIX Conference on Security Symposium*, SEC'14, pages 575–590, Berkeley, CA, USA, 2014. USENIX Association.

[53] A. Forget, S. Chiasson, P. C. van Oorschot, and R. Biddle. Improving text passwords through persuasion. In *Proceedings of the 4th Symposium on Usable Privacy and Security*, SOUPS '08, pages 1–12, New York, NY, USA, 2008. ACM.

[54] M. Frank, R. Biedert, E.-D. Ma, I. Martinovic, and D. Song. Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication. *Information Forensics and Security, IEEE Transactions on*, pages 136–148, 2013.

[55] S. Gaw and E. W. Felten. Password management strategies for online accounts. In *Proceedings of the Second Symposium on Usable Privacy and Security*, SOUPS '06, pages 44–55, New York, NY, USA, 2006. ACM.

[56] S. Gooding. Ridiculously smart password meter coming to wordpress 3.7. Accessed: 2016-04-01 from `http://wptavern.com/ridiculously-smart-password-meter-coming-to-wordpress-3-7`.

[57] Google. Google ngram viewer, 2013. Retrieved May 3 2015 from `http://storage.googleapis.com/books/ngrams/books/datasetsv2.html`.

[58] B. Grawemeyer and H. Johnson. Using and managing multiple passwords: A week to a view. *Interact. Comput.*, 23(3), 2011.

[59] K. Greene, M. Gallagher, B. Stanton, and P. Lee. I Can't Type That! P@$$w0rd Entry on Mobile Devices. In *Human Aspects of Information Security, Privacy, and Trust*. Springer International Publishing, 2014.

[60] S. M. T. Haque, M. Wright, and S. Scielzo. Passwords and interfaces: Towards creating stronger passwords by using mobile phone handsets. In *Proc. of SPSM '13*, 2013.

[61] S. T. Haque, M. Wright, and S. Scielzo. Hierarchy of users web passwords: Perceptions, practices and susceptibilities. *International Journal of Human-Computer Studies*, 72(12):860–874, 2014.

[62] M. Harbach, E. von Zezschwitz, A. Fichtner, A. D. Luca, and M. Smith. It's a hard lock life: A field study of smartphone (un)locking behavior and risk perception. In *Proc. SOUPS' 14*, pages 213–230, Menlo Park, CA, 2014. USENIX Association.

[63] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183, 1988.

[64] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, pages 139–183, 1988.

[65] R. V. Hartley. Transmission of information. *Bell System technical journal*, 7(3):535–563, 1928.

[66] hashcat. hashcat.

[67] hashcat. oclhashcat - advanced password recovery, 2015. Retrieved May 2 2015 from `http://hashcat.net/oclhashcat/`.

[68] E. Hayashi and J. Hong. A diary study of password usage in daily life. In *Proc. CHI '11*, pages 2627–2630. ACM, 2011.

[69] C. Herley and P. van Oorschot. A research agenda acknowledging the persistence of passwords. *IEEE Security and Privacy*, 2012.

[70] J. H. Huh, H. Kim, R. B. Bobba, M. N. Bashir, and K. Beznosov. On the memorability of system-generated pins: Can chunking help? In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 197–209, Ottawa, July 2015. USENIX Association.

[71] K. Inc. Korelogic security, 2015. Retrieved May 2 2015 from `http://www.korelogic.com`.

[72] P. G. Inglesant and M. A. Sasse. The true cost of unusable password policies: Password use in the wild. In *Proc. CHI '10*, pages 383–392. ACM, 2010.

[73] M. Jakobsson and R. Akavipat. Rethinking passwords to adapt to constrained keyboards. In *Proc. of MoST*, 2012.

[74] A. Johnson. The speed of mental rotation as a function of problem-solving strategies. *Perceptual and Motor Skills, 71, 803-806.*, 1990.

[75] A. Juels and M. Sudan. A fuzzy vault scheme. *Designs, Codes and Cryptography*, 38(2):237–257, 2006.

[76] M. J. Kahana. *Foundations of human memory.* Oxford University Press, 2012.

[77] P. G. Kelley, S. Komanduri, M. L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, and J. Lopez. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *2012 IEEE Symposium on Security and Privacy*, pages 523–537, May 2012.

[78] P. G. Kelley, S. Komanduri, M. L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, and J. Lopez. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *Proc. of SS&P'12*, 2012.

[79] J. F. Kihlstrom. Memory research: The convergence of theory and practice. *Basic and Applied Memory Research: Volume 1: Theory in Context; Volume 2: Practical Applications*, page 5, 2013.

[80] S. Komanduri, R. Shay, L. F. Cranor, C. Herley, and S. Schechter. Telepathwords: Preventing weak passwords by reading users' minds. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 591–606, San Diego, CA, Aug. 2014. USENIX Association.

[81] S. Komanduri, R. Shay, P. G. Kelley, M. L. Mazurek, L. Bauer, N. Christin, L. F. Cranor, and S. Egelman. Of passwords and people: Measuring the effect of password-composition policies. In *Proceedings of the SIGCHI Conference on*

Human Factors in Computing Systems, CHI '11, pages 2595–2604, New York, NY, USA, 2011. ACM.

[82] KoreLogic. Korelogic hashcat rules, 2010. Retrieved May 1 2015 from `http://contest-2010.korelogic.com/rules-hashcat.html`.

[83] J. Lawler. The "web2" file of english words, 1999. Retrieved May 3 2015 `http://www-personal.umich.edu/~jlawler/wordlist`.

[84] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[85] Y. Li. Protractor: A fast and accurate gesture recognizer. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 2169–2172, 2010.

[86] Z. Li, W. Han, and W. Xu. A large-scale empirical analysis of chinese web passwords. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 559–574, San Diego, CA, Aug. 2014. USENIX Association.

[87] I. S. MacKenzie and K. Tanaka-Ishii. *Text Entry Systems: Mobility, Accessibility, Universality*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2007.

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

[89] J. A. McGeoch. Forgetting and the law of disuse. *Psychological review*, 39(4):352, 1932.

[90] N. Micallef, M. Just, L. Baillie, M. Halvey, and H. G. Kayacik. 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, pages 284–294, 2015.

[91] W. Moncur and G. Leplâtre. Pictures at the atm: Exploring the usability of multiple graphical passwords. In *Proc. CHI '07*, pages 887–894, New York, NY, USA, 2007. ACM.

[92] M. A. Nacenta, Y. Kamber, Y. Qiang, and P. O. Kristensson. Memorability of pre-designed and user-defined gesture sets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1099–1108, 2013.

[93] J. Nicholson, L. Coventry, and P. Briggs. Age-related performance issues for pin and face-based authentication systems. In *Proc. CHI '13*, pages 323–332. ACM, 2013.

[94] G. Notoatmodjo and C. Thomborson. Passwords and perceptions. In *Proceedings of the Seventh Australasian Conference on Information Security - Volume 98*, AISC '09, pages 71–78, Darlinghurst, Australia, Australia, 2009. Australian Computer Society, Inc.

[95] U. Oh and L. Findlater. The challenges and potential of end-user gesture customization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1129–1138, New York, NY, USA, 2013. ACM.

[96] Openwall. John the ripper password cracker. Retrieved August 24 2015 from `http://www.openwall.com/john/`.

[97] Openwall. John the ripper single crack mode. Retrieved August 24 2015 from `http://www.openwall.com/john/doc/MODES.shtml`.

[98] Openwall. Openwall wordlists collection. Retrieved Auguts 26 2015 from `http://www.openwall.com/wordlists/`.

[99] A. Oulasvirta, A. Reichel, W. Li, Y. Zhang, M. Bachynskyi, K. Vertanen, and P. O. Kristensson. Improving two-thumb text entry on touchscreen devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2765–2774, 2013.

[100] A. Oulasvirta, T. Roos, A. Modig, and L. Leppänen. Information capacity of full-body movements. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1289–1298, New York, NY, USA, 2013. ACM.

[101] A. Oulasvirta, S. Tamminen, V. Roto, and J. Kuorelahti. Interaction in 4-second bursts: The fragmented nature of attentional resources in mobile hci. In *Proc. CHI '05*, pages 919–928. ACM, 2005.

[102] outpost9. A list of popular password caracking wordlist, 2005. Retrieved August 27 2015 from `http://www.outpost9.com/files`.

[103] A. Paivio and K. Csapo. Picture superiority in free recall: Imagery or dual coding? *Cognitive psychology*, 5:176–206.

[104] B. Poppinga, A. Sahami Shirazi, N. Henze, W. Heuten, and S. Boll. Understanding shortcut gestures on mobile touch devices. In *Proc. MobileHCI '14*, pages 173–182. ACM, 2014.

[105] N. Sae-Bae, K. Ahmed, K. Isbister, and N. Memon. Biometric-rich gestures: A novel approach to authentication on multi-touch devices. In *Proc. CHI '12*, pages 977–986. ACM, 2012.

[106] M. Sasse, M. Steves, K. Krol, and D. Chisnell. The great authentication fatigue and how to overcome it. In *Cross-Cultural Design*, Lecture Notes in Computer Science, pages 228–239. Springer International Publishing, 2014.

[107] F. Schaub, R. Deyhle, and M. Weber. Password entry usability and shoulder surfing susceptibility on different smartphone platforms. In *Proc. of MUM '12*, 2012.

[108] S. Schechter, C. Herley, and M. Mitzenmacher. Popularity is everything: A new approach to protecting passwords from statistical-guessing attacks. In *Proceedings of the 5th USENIX Conference on Hot Topics in Security*, HotSec'10, pages 1–8, Berkeley, CA, USA, 2010. USENIX Association.

[109] S. Schneegass, F. Steimle, A. Bulling, F. Alt, and A. Schmidt. Smudgesafe: Geometric image transformations for smudge-resistant user authentication. In

Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '14, pages 775–786, New York, NY, USA, 2014. ACM.

[110] K. F. Schulz and D. A. Grimes. Unequal group sizes in randomised trials: guarding against guessing. *The Lancet*, 2002.

[111] SCOWL. Spell checker oriented word lists (and friends). Retrieved Auguts 27 2015 from `http://wordlist.sourceforge.net`.

[112] A. Serwadda and V. V. Phoha. When kids' toys breach mobile phone security. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, CCS '13, pages 599–610, New York, NY, USA, 2013. ACM.

[113] M. Shahzad, A. X. Liu, and A. Samuel. 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, pages 39–50, 2013.

[114] C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 2001.

[115] R. Shay, L. Bauer, N. Christin, L. F. Cranor, A. Forget, S. Komanduri, M. L. Mazurek, W. Melicher, S. M. Segreti, and B. Ur. A spoonful of sugar?: The impact of guidance and feedback on password-creation behavior. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2903–2912, New York, NY, USA, 2015. ACM.

[116] R. Shay, P. G. Kelley, S. Komanduri, M. L. Mazurek, B. Ur, T. Vidas, L. Bauer, N. Christin, and L. F. Cranor. Correct horse battery staple: Exploring the usability of system-assigned passphrases. In *Proc. of SOUPS '12*, 2012.

[117] R. Shay, S. Komanduri, A. L. Durity, P. S. Huh, M. L. Mazurek, S. M. Segreti, B. Ur, L. Bauer, N. Christin, and L. F. Cranor. Can long passwords be secure and usable? In *Proc. of CHI '14*, 2014.

[118] M. Sherman, G. Clark, Y. Yang, S. Sugrim, A. Modig, J. Lindqvist, A. Oulasvirta, and T. Roos. 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, pages 176–189, 2014.

[119] D. Silver, S. Jana, D. Boneh, E. Chen, and C. Jackson. Password managers: Attacks and defenses. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 449–464, San Diego, CA, Aug. 2014. USENIX Association.

[120] SkullSecurity. Return of the facebook snatchers, 2010. Retrieved August 24 2015 from `https://blog.skullsecurity.org/2010/return-of-the-facebook-snatchers`.

[121] A. Smith. U.s. smartphone use in 2015, 2015. Retrieved August 27 2015 from `http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/`.

[122] Y. Song, G. Cho, S. Oh, H. Kim, and J. H. Huh. 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, pages 2343–2352, 2015.

[123] T. SpiderLabs. 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-/`.

[124] SRLab. Spoofing fingerprints, 2013. Retrieved Feb 12 2015 from `https://srlabs.de/spoofing-fingerprints/`.

[125] E. Stobert and R. Biddle. Memory retrieval and graphical passwords. In *Proc. SOUPS '13*, pages 15:1–15:14. ACM, 2013.

[126] E. L. Thorndike. *The psychology of learning*, volume 2. Teachers College, Columbia University, 1913.

[127] J. Tian, C. Qu, W. Xu, and S. Wang. Kinwrite: Handwriting-based authentication using kinect. In *Proceedings of NDSS 2013*, 2013.

[128] S. Uellenbeck, M. Dürmuth, C. Wolf, and T. Holz. 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, pages 161–172, 2013.

[129] B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. L. Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, N. Christin, and L. F. Cranor. How does your password measure up? the effect of strength meters on password creation. In *Proc. Security'12*, pages 5–5. USENIX Association, 2012.

[130] B. Ur, S. M. Segreti, L. Bauer, N. Christin, L. F. Cranor, S. Komanduri, D. Kurilova, M. L. Mazurek, W. Melicher, and R. Shay. Measuring real-world accuracies and biases in modeling password guessability. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 463–481, Aug. 2015.

[131] R. Veras, C. Collins, and J. Thorpe. On semantic patterns of passwords and their security impact. In *NDSS 2014*, 2014.

[132] E. von Zezschwitz, A. De Luca, and H. Hussmann. *Survival of the Shortest: A Retrospective Analysis of Influencing Factors on Password Composition*, pages 460–467. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[133] E. von Zezschwitz, A. De Luca, P. Janssen, and H. Hussmann. 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, pages 2339–2342, 2015.

[134] E. von Zezschwitz, P. Dunphy, and A. De Luca. Patterns in the wild: A field study of the usability of pattern and pin-based authentication on mobile devices. In *Proc. MobileHCI '13*, pages 261–270. ACM, 2013.

[135] K.-P. L. Vu, R. W. Proctor, A. Bhargav-Spantzel, B.-L. B. Tai, J. Cook, and E. Eugene Schultz. Improving password security and memorability to protect personal and organizational information. *Int. J. Hum.-Comput. Stud.*, 65(8):744–757, Aug. 2007.

[136] M. Weir. Optimizing john the ripper's "single" mode for dictionary attacks, 2010. Retrieved October 4 2014 from `http://reusablesec.blogspot.com/2010/04/optimizing-john-rippers-single-mode-for.html`.

[137] M. Weir, S. Aggarwal, M. Collins, and H. Stern. Testing metrics for password creation policies by attacking large sets of revealed passwords. In *Proc. of CCS'10*, 2010.

[138] M. Weir, S. Aggarwal, B. d. Medeiros, and B. Glodek. Password cracking using probabilistic context-free grammars. In *Proc. of SP '09*, 2009.

[139] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, pages 80–83, 1945.

[140] J. Yan, A. Blackwell, R. Anderson, and A. Grant. Password memorability and security: Empirical results. *IEEE Security and Privacy*, 2(5):25–31, Sept. 2004.

[141] S. Zhai, M. Hunter, and B. A. Smith. Performance optimization of virtual keyboards. *Human–Computer Interaction*, 2002.

[142] Y. Zhang, F. Monrose, and M. K. Reiter. The security of modern password expiration: an algorithmic framework and empirical analysis. In *Proc. of CCS '10*.

[143] N. Zheng, K. Bai, H. Huang, and H. Wang. You are how you touch: User verification on smartphones via tapping behaviors. In *Network Protocols (ICNP), 2014 IEEE 22nd International Conference on*, pages 221–232, 2014.

Appendix A

Memorability study

A.1 Application Screenshots

Figure A.1 and A.2 shows the sample screenshots of the web application we used in the study.

A.2 Accounts

We have eight accounts, four were designed to be simple accounts, and other four were strong accounts. The naming and description of the accounts were shown in Table A.1.

Name	Type	Description
Pacificx Bank	Strong	Online Banking account allows you to view your balance, withdraw or deposit cash.
YaMail	Strong	YaMail is the new email vendor that allows you to communicate easier and faster.
DealsMoon	Simple	DealsMoon has all the hot and popular deals that save you money!
FaceNote	Strong	THE world NO.1 social network site that let everybody knows what you are doing.
Artsy eCommerce	Strong	Start your business today.
Trut Weather	Simple	Provide you the daily-updated accurate weather of your place.
Old-fashion news	Simple	Your daily source of news.
Hifi Music	Simple	Streaming the music of best quality to everybody.

Table A.1: The account information of our memorability study. For some accounts we used the description to help participants better understand the type of the acocunt.

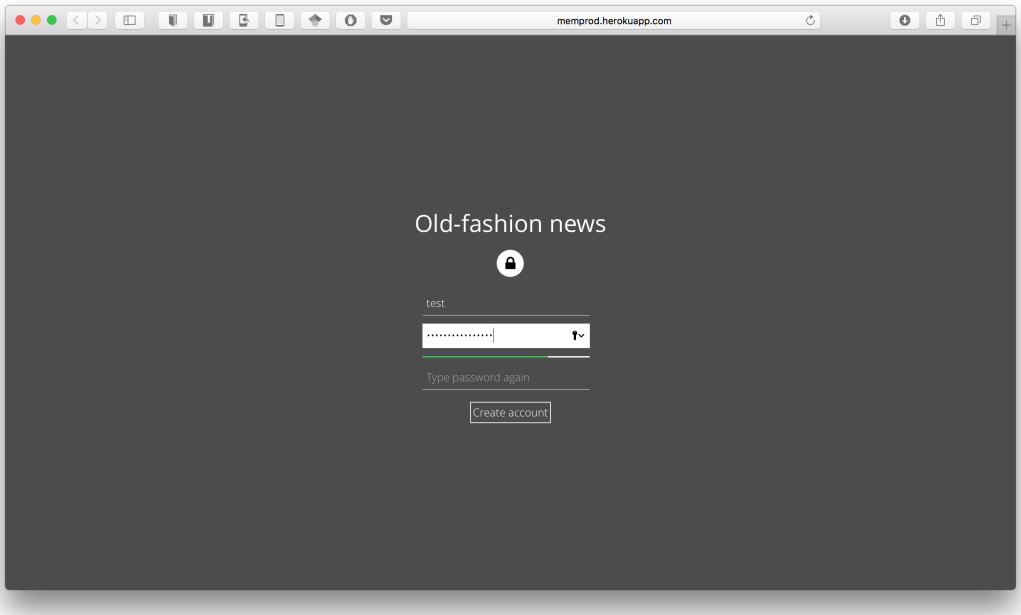


Figure A.1: A screenshot of our web application. It shows the creation interface for one of the accounts.

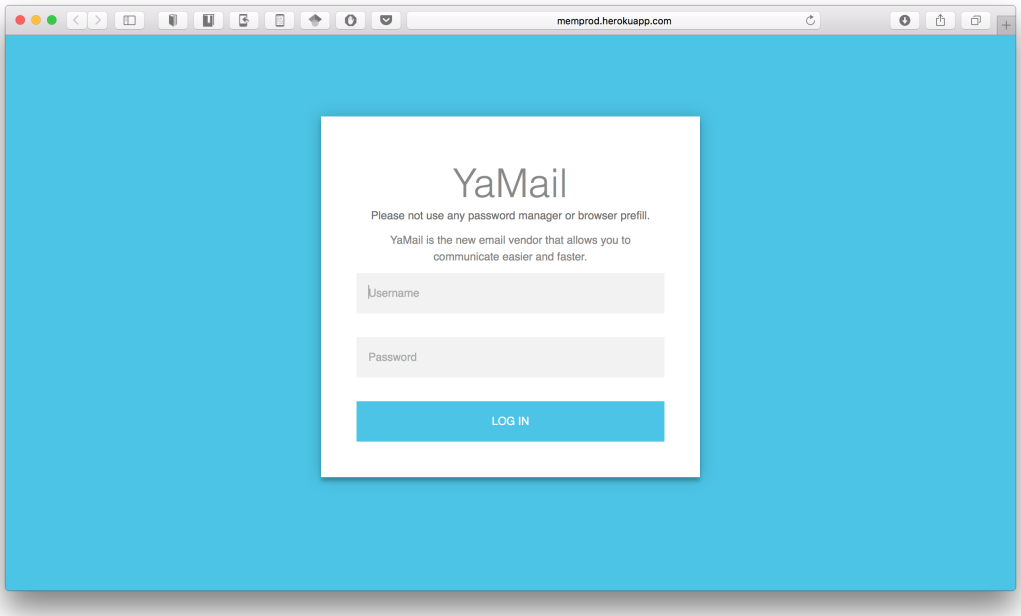


Figure A.2: A screenshot of our web application. It shows the log-in interface for one of the accounts.