

HUMAN FACTORS IN TEXTUAL PASSWORD-BASED AUTHENTICATION

by

S M TAIABUL HAQUE

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

August 2015

Copyright © by S M Taiabul Haque 2015

All Rights Reserved

To my mother and my father
who guided me to where I am today.

ACKNOWLEDGEMENTS

I would like to express my profound gratitude to Dr. Matthew Wright for his generous guidance, constant motivation, and invaluable support as my supervising professor. I am deeply grateful to my dissertation committee member Dr. Shannon Scielzo for actively helping me with my research work and arranging the participant pool for user studies. I wish to thank my other committee members Dr. Donggang Liu, and Dr. Vassilis Athitsos for their interest in my research and for their time in reviewing my work.

I so greatly appreciate the financial support provided by the graduate school of UT Arlington throughout my doctoral studies. I am thankful to all my teachers in Bangladesh and United States who made me what I am today.

Finally, I would like to express my deepest gratitude to my parents who have encouraged me to pursue my PhD in United States. I also thank my friends and fellow members of the iSec lab at UTA for their helpful comments and suggestions.

July 10, 2015

ABSTRACT

HUMAN FACTORS IN TEXTUAL PASSWORD-BASED AUTHENTICATION

S M Taiabul Haque, Ph.D.

The University of Texas at Arlington, 2015

Supervising Professor: Matthew Wright

Despite being the most commonly used method of authentication on the Web, textual password-based authentication is by no means a panacea as long as usability is concerned. In this dissertation work, we address some usability issues of textual password-based authentication and propose solutions to them. In our first work, we propose a hierarchy of password importance and use an experiment to examine the degree of similarity between passwords for lower-level (e.g. news portal) and higher-level (e.g. banking) websites in this hierarchy. Leveraging the lower-level passwords constructed by subjects along with a password-cracking dictionary, we successfully cracked almost one-third of the subjects' higher-level passwords. This confirms that leaked lower-level passwords can be used by attackers to crack higher-level passwords.

In our second work, we examine the issue of textual password entry on mobile devices which is fraught with usability problems due to size and input constraints of mobile devices. We examine the association between password strengths and the keyboard/keypad layouts through which they are constructed, including computer keyboard and different types of mobile keypad layouts. We design a custom mobile keypad layout and demonstrate its effectiveness through extensive user studies.

Our third work focuses on measuring user comfort when constructing a strong password by using mobile devices. Since comfort is a basic construct for understanding usability, measuring user comfort in a security context is an issue of paramount importance. We solve this issue by applying standard techniques of psychometrics to develop a user comfort scale. We establish the essential psychometric properties (reliability and validity) of this scale and demonstrate how the scale can be used to profile password construction interfaces of popular smartphone handsets. We also theoretically conceptualize user comfort across different dimensions and use confirmatory factor analysis to verify our theory.

All these works reveal the weaknesses of user-chosen textual passwords. Thus, in our final work, we focus on system-assigned random textual password consisting of lowercase letters only. It guards against a wide range of usability issues, but introduces memorability problem, which hinders its wide-scale deployment in real world. We propose two methods to leverage different types of human memory and aid the users in memorizing system-assigned random passwords. The first method (known as the method of loci) exploits the spatial and the visual memory to help memorizing a list of ordered items. The second method (known as the link method), on the other hand, facilitates the memorization process by creating a chain of memory cues. We implemented both of the methods in the context of memorizing system-assigned random passwords and conducted a memorability study to test their effectiveness. We found that participants using the method of loci had a login success rate of 86%, which is highest for any recall-based study with system-assigned random passwords. By extending the method of loci, we further conducted a separate study to test its effectiveness in helping users to memorize long random passwords that offer almost crypto-level security. The results of this study demonstrate that the method of loci can be leveraged to help users memorize cryptographically-strong passwords.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iv
ABSTRACT	v
LIST OF ILLUSTRATIONS	xii
LIST OF TABLES	xiii
Chapter	Page
1. Introduction	1
2. Related Work	9
2.1 User-chosen Passwords	9
2.2 Password Reuse and Password Similarity	10
2.3 Mobile Device Password Entry	14
2.4 Questionnaire Development	15
2.4.1 Usability Questionnaires in HCI	15
2.4.2 Psychometrics in HCI	16
2.4.3 Psychological Approach in Usable Security	17
2.5 System-assigned Random Passwords	17
3. Hierarchy of Users' Web Passwords: Perceptions, Practices, and Suscepti- bilities	18
3.1 Hierarchy	20
3.2 Hypotheses and Research Questions	22
3.3 Methodology	25
3.3.1 Study Administration	26
3.3.2 First Phase	26

3.3.3	Second Phase	31
3.3.4	Survey Analysis	33
3.4	Results	34
3.4.1	Demographics	34
3.4.2	Perceived Importance of Passwords	35
3.4.3	Sketchy Password	41
3.4.4	Content Password	44
3.4.5	Password Sharing	45
3.4.6	Password Cracking	47
3.5	Discussion	52
3.5.1	Limitations	52
3.5.2	Implications	54
3.5.3	Conclusion	57
4.	Towards Creating Stronger Passwords by Using Mobile Phone Handsets . .	59
4.1	Experiment 1	59
4.1.1	Study Administration	59
4.1.2	Apparatus	59
4.1.3	Experimental Groups	60
4.1.4	Password Construction	61
4.1.5	Results	62
4.1.6	Discussion	64
4.2	Experiment 2	66
4.2.1	Password Construction	66
4.2.2	Results	68
4.2.3	Discussion	69
4.3	General Discussion	71

4.3.1	Ecological Validity	71
4.3.2	Low Sample Size	71
4.3.3	Standard Custom Layout	72
4.3.4	Usability Evaluation	72
4.3.5	Password Strength Measurement	72
5.	Applying Psychometrics to Measure User Comfort when Constructing a Strong Password	73
5.1	Psychometrics	73
5.1.1	Reliability and Validity	74
5.1.2	Framework	75
5.2	Scale Development Steps	76
5.2.1	Domain Specification and Initial Item Pool Generation	76
5.2.2	Content Validity Assessment	79
5.2.3	Initial Scale Administration – Study 1	82
5.2.4	Reliability Analysis	85
5.2.5	Construct Validity Assessment	86
5.2.6	Criterion Validity Assessment – Study 2	87
5.3	Profiling Popular Smartphone Handset Interfaces – Study 3	91
5.3.1	Participants	92
5.3.2	Procedure	92
5.3.3	Results	92
5.4	Factor Analysis	93
5.5	Discussion	96
5.5.1	Ecological validity	97
5.5.2	Limitations	98
5.5.3	Aggregation and application	99

5.5.4	Norm development	99
5.5.5	Revalidation study	100
6.	Novel Techniques for Memorizing System Assigned Random Passwords . .	102
6.1	Constraints and Objectives	102
6.1.1	Entropy	102
6.1.2	Time	103
6.1.3	Automation	104
6.2	The Memory Techniques	104
6.2.1	The Science Behind Proposed Techniques	105
6.3	System Design	109
6.3.1	Method of Loci	109
6.3.2	Link Method	111
6.3.3	Pilot Study	112
6.3.4	Development Platform and Tools	112
6.4	User Study	113
6.4.1	Participants, Apparatus and Environment	113
6.4.2	Procedure	114
6.4.3	Ecological Validity	115
6.5	Results	115
6.5.1	Significance Tests	116
6.5.2	Memorability	117
6.5.3	Registration Time	119
6.5.4	Number of Attempts	119
6.5.5	Login Time	120
6.6	System Analysis	120

6.7	Memorizing Cryptographically-Strong Passwords by Leveraging the Method of Loci	126
6.7.1	Cryptographically-Strong Passwords	126
6.7.2	Spaced Repetition for Memorizing Cryptographically-Strong Pass- words	127
6.8	Method of Loci for Memorizing Cryptographically-Strong Passwords .	128
6.8.1	Participants, Apparatus, Environment and Procedure	128
6.8.2	Results	130
6.9	Conclusion	131
	REFERENCES	133
	BIOGRAPHICAL STATEMENT	146

LIST OF ILLUSTRATIONS

Figure	Page
3.1 User password hierarchy based on the perceived importance level. . .	23
3.2 Mock password construction page for Weather.com.	29
3.3 Mock password construction page for Dreamdeals.com.	30
3.4 Box plot of participant ratings on a 1 to 5 scale (1 being “not impor- tant”, 3 being “moderately important”, 5 being “important”) about the perceived importance of passwords of different websites.	36
3.5 A comparison of mean lengths.	39
3.6 A comparison among passwords of different categories.	40
3.7 Password cracking statistics (without wordlist).	49
3.8 Password reuse mechanism.	50
3.9 Password cracking statistics (with wordlist).	51
4.1 Custom layout with two extra rows of keys	60
4.2 Box plot of entropy values. The mean values have been highlighted by the black points.	63
6.1 Atkinson-Shiffrin Memory Model	105
6.2 Login success rates for the study conditions [N = 44]	116
6.3 Login time for the study conditions	120

LIST OF TABLES

Table	Page
3.1 Summary of post hoc analysis with Wilcoxon signed-rank tests with a Bonferroni correction applied. For brevity, all entries with $p < .00179$ have been omitted.	37
3.2 Type of modifications made when reusing an existing password to create a sketchy password.	42
3.3 Type of similarity between an existing password and the new sketchy password.	43
3.4 Type of modifications made when reusing an existing password to create a shared password.	46
3.5 Type of similarity between an existing password and the new shared password.	47
4.1 Summary of Tukey's post-hoc analysis. For each pair of interfaces, the difference, the 95% confidence interval and the p-value of the pairwise comparison are shown.	64
5.1 Reliability Analysis. Cronbach's α value is 0.96.	81
6.1 Number of Attempts for Successful Logins [SD: Standard Deviation] .	118

CHAPTER 1

Introduction

More than a decade and a half ago, the seminal paper on usable security, titled as “Why Johnny Can’t Encrypt”, was first published where the authors demonstrated that PGP 5.0, the most popular encryption software of that period, was essentially unusable by average computer users, mainly because of its user interface design flaws [119]. Since then, security researchers began to acknowledge the importance of studying users to identify the underlying human factors, and designing a security system accordingly [109, 103]. Authentication, being an important subfield of security, received much attention among usable security researchers. Many authentication schemes have been proposed to search a balance between usability and security, ranging from drawing a doodle to identifying a spot on Google Maps [38, 45, 115]. However, none could beat the simplicity and cost effectiveness of typing a sequence of characters as the authentication secret [11]. As a result, textual password-based authentication scheme still remains the most favorable form of user authentication on the Web, and there is little probability that the scenario might change in the near future [55].

Unfortunately, textual password-based authentication is by no means a panacea as long as usability is concerned. As formulated by Wiedenbeck et al., a good password needs to satisfy two conflicting requirements at the same time: being “easy to remember” and “hard to guess” [120]. Naturally, passwords that are easy to remember are short single words found in dictionaries or wordlists, or slight variations. Choosing such words as passwords makes them vulnerable to dictionary attacks. Personally

meaningful words or numbers are also memorable, but they are easily guessable as well, especially by friends or family members.

Many studies have been conducted for understanding the password habits of users. Researchers from Microsoft Research did a landmark study that involved half a million users and revealed many interesting findings about different user password habits [36]. Academic researchers, on the other hand, used various novel methods and laboratory studies to observe more closely a particular password behavior (e.g., password reuse habit) of a sample population [107, 41]. Although these papers have reported about password reuse, we are aware of no such work in the literature which looked at finer-grained aspects of reuse – how similar are passwords to one another across sites, and how do they vary with perceived security level. Our first work attempted to fill this gap in the existing literature.

The work was inspired by the findings of Notoatmodjo et al., which confirmed that users mentally group their accounts and tend to make stronger passwords for accounts that they consider more important [83]. Users have different levels of incentive to protect their different accounts. In this work, we examined how vulnerable the higher-level (webmail or banking account) passwords of a specific user would become, if the lower-level (online news account or weather portal) passwords of that user could be compromised. Our results showed that, almost one-third of the higher-level passwords of the participants could be cracked by using the lower-level passwords and a comprehensive wordlist. This demonstrated that, the knowledge of a password of a lower risk account seems to increase the chance to crack higher risk account based on similarity to the lower risk password.

The password construction problem has become more aggravated with the proliferation of mobile devices. Password entry is a time-consuming and error-prone operation on mobile handsets. A study by Jakobsson et al. reported that password

entry on handsets frustrates users even more than lack of coverage, small screen size, or poor voice quality on such devices [60].

This poor user experience raises an important research question: “How do input-constrained devices like mobile phone handsets affect the password behavior of users?”. In general, security experts say that a good password includes a combination of uppercase and lowercase letters, digits, and special characters¹. Capitalizing a lowercase letter or inserting a digit on a mobile phone handset is not as straightforward as it is on a computer keyboard. On an iPhone, for example, each shift to and from digits requires one extra click. On the other hand, a digit can be inserted in the same way as a letter on a computer keyboard with no extra effort. Since the auto-correction and auto-completion options of mobile handsets are not enabled for the password field, the general comfort of typing a password is also less on a handset than a computer.

These limitations suggest that passwords that are constructed by using mobile handsets would be relatively weaker than those constructed by using computer keyboards. However, there exists no empirical work in the current literature that examines the association between the strength of a password and the interface through which it is constructed. Moreover, mobile handsets can also be classified into two categories: handsets with physical keyboards and handsets with touchscreen keypads. These two layouts present the users two different experiences when typing.

The goal of our second work was to examine how password strengths vary with the keyboard or keypad layout through which they are constructed. We were also interested to observe the behavior of users when they are presented with a more convenient interface to construct passwords on a mobile handset. Therefore, we designed and evaluated the effectiveness of a custom layout that presented the users a

¹See, e.g., <http://www.us-cert.gov/ncas/tips/st04-002>

more convenient option to insert digits and special characters when constructing a password. The user study results confirmed that if users are presented with a more convenient method of entering digits and special characters on mobile handsets, they take advantage of it to construct stronger passwords.

The study results also highlighted that users are not comfortable with the existing textual password entry mechanisms on mobile phone handsets. This poor user experience clearly undermines the usability of sensitive security systems that are developed for mobile platforms (mobile banking, for example). According to Whitten and Tygar in their seminal paper, a security system is deemed to be usable if “people are sufficiently comfortable with the interface to continue using it” [119]. In a mobile banking system, a user is required to type her entire password by using the mobile phone keypad (i.e. no “remember me” option) each time she intends to log in to her bank account. Thus, user frustrations on password entry on mobile handsets could undermine the usability of mobile banking as a whole, as well as other security systems on mobile devices.

Since “frustration” and “comfort” are subjective psychological concepts, it is not a straightforward task to measure the level of comfort a user feels when using the interface of a security system. According to psychology researchers [86, 78, 111], merely asking “How much comfortable you are with the interface of this security system?” is not enough in this case for three reasons. First, a single question lacks scope to represent a complex psychological concept such as comfort. Just as a single question can not measure intelligence, a single question is not sufficient for measuring one’s level of comfort. Second, a single question can only categorize people into a small number of groups, thus limiting the ability to discriminate among finer degrees. Third, any individual question has a considerable amount of measurement error associated

with it. When multiple questions are asked and the response scores are summed to get a total score, this error tends to average out.

For these reasons, to measure complex psychological concepts such as “frustration” or “comfort”, psychology researchers develop a set of questions that meets some widely agreed upon specific statistical criteria. In fact, a separate branch of psychology has evolved in this regard, which is known as *psychometrics*. Psychometrics concentrates on developing and validating questionnaires or tests that are used for assessing knowledge, attitudes, abilities, or various personality traits.

In our third work, we adopted the methods of psychometrics to develop a questionnaire, also called a scale, for measuring the comfort of constructing a strong password when using a particular interface. We first used expert opinions to guide the creation and selection of questions and then assessed our questionnaire for reliability and validity, the two essential psychometric properties of a scale. To this end, we conducted two user studies where we administered the questionnaire to undergraduate students from different majors and analyzed their responses. We found that our questionnaire meets all of the requirements for reliability and validity for a psychometric scale: it is consistent, complete, accurately focused, and capable of predicting certain real-world outcomes.

Through a separate user study, we evaluated the password construction interfaces of popular smartphone handsets by using our scale, where the interface of iPhone was rated the most comfortable by the participants. The results of these studies demonstrated that our scale can be used effectively to measure user comfort during a password entry operation on a mobile handset.

Based on certain observations, we further shortened our scale while maintaining the diversity of interface quality evaluation. We hypothesized a specific theory about user comfort in constructing a strong password by conceptualizing comfort across

several factors and built a four-factor model. This model is helpful to explain why a particular interface is more comfortable to use than another one. We employed confirmatory factor analysis, a widely used statistical method in psychometrics, and found that our collected user responses fit the model we developed.

To the best of our knowledge, despite being a well-developed field, psychometrics has not been applied in usable security to develop and evaluate questionnaires. We believe that our work paved the way for applying the techniques of psychometrics in measuring various subjective concepts that are associated with usable security. In particular, our psychometric approach of measuring comfort can be generalized to measure whether people are sufficiently comfortable with other security-related interfaces, such as anti-virus systems, personal firewalls, privacy tools for the Web, and encryption software. This, in turn, would be helpful to understand in what ways that interface is usable or not, according to the working definition of usable security as provided by Whitten and Tygar [119].

While our first work demonstrated the consequences of password reuse across websites of different importance levels, the results from our second and third works highlighted the inconvenience of entering uppercase letters, digits, and special characters when constructing a password by using a mobile device. Thus, we next focused on system-assigned textual passwords consisting of lowercase letters only, in which the system randomly generates a sequence of lowercase letters for a user to be used as her password. Since each system randomly generates its own password, password reuse across websites is precluded. Moreover, lowercase letter only passwords increase the convenience of password entry when using mobile devices.

System-assigned textual password also guards against a wide range of usability issues. Since the password is assigned randomly, the susceptibility of dictionary or guessing attacks is removed. Furthermore, it guards against poor password choices

from users. The major bottleneck, on the other hand, is the memorability problem. For a user, a random sequence of characters is not as memorable as a plain dictionary word or her mother’s maiden name. This might tend the user to write down the password which is considered to be a bad security practice [56, 130].

Numerous study results have confirmed the poor memorability issue for system-assigned random passwords [124, 129]. A separate study has assessed the memorability of system assigned passphrases (space-delimited sets of natural language words), but the results were not encouraging and did not show significant improvement over system-assigned textual password of similar entropy [106].

In our final work, we addressed this crucial issue of memorability for system-assigned random passwords. We proposed novel methods to leverage different types of human memory and aid the users in memorizing system-assigned random passwords in an effective way. We reviewed the literature of memory and memorization techniques and picked two scientifically proven methods that would be pertinent to password memorization. The first method (known as the method of loci or the memory palace method) exploits the spatial and the visual memory to help memorizing a list of ordered items. The second method (known as the link method or story method), on the other hand, facilitates the memorization process by creating a chain of memory cues.

We implemented both of the methods in the context of memorizing system-assigned random textual passwords and conducted a two-part memorability study with 52 users to test their effectiveness. Our study results showed that both of the methods provided significant improvement over the control condition with regard to rate of unsuccessful logins due to memory recall failure. In particular, the method of loci had a login success rate of 86% within three attempts, which is highest for any recall-based study with system-assigned random passwords. With a registration time

of 160 seconds and a median login time of 9 seconds, the method of loci offers a good solution to the usability-security tradeoff in the field of user authentication.

We further extended the method of loci to observe its effectiveness in assisting users to memorize passwords that offer almost crypto-level security [9]. Unlike previous work, we did not relax the time constraint too much in this regard [13]. The results of our two-part memorability study showed that the method of loci could be leveraged to help users memorize cryptographically-strong password in just a single session. We believe that our adoption of the method of loci in the context of password memorization is a major step forward in solving the dilemma of memorability and guessability in textual password-based authentication mechanism.

CHAPTER 2

Related Work

Prior researchers have identified the usability problems of user-chosen passwords, deeply investigated password reuse and similarity issues, proposed novel methods for mobile device password entry, developed usability questionnaires, and tried to improve the memorability of system-assigned random passwords. We now discuss these works.

2.1 User-chosen Passwords

Study results have verified that when users are given the freedom of choosing their own passwords, they create weak passwords that are easy to remember and contain predictable patterns [1, 128, 10]. This, in turn, makes the passwords vulnerable to dictionary attacks. Although strict password composition policies prevent users from creating weak passwords, these policies might lead to user frustration at times without providing considerable security benefit [107, 68]. Shay et al. confirm that a strict password composition policy has an adverse effect on memorability [107]

Another approach is to ask users to create their own passwords, but proactively check those password strengths by deploying a password meter [10, 19, 117]. However, a recent comprehensive study on password meters demonstrates the weaknesses and inconsistencies with regard to the deployment of password meters on real-world websites [31]. This suggests that password researchers have yet to reach a consensus regarding the strengths of various user-chosen passwords.

Password manager is another tool which helps users in storing and using multiple strong passwords that are generated with complex password policies. Unfortunately, numerous security analyses have demonstrated that password managers are vulnerable to various types of attack [112, 40, 127]. As a result, password managers have yet to provide a definitive solution to the password management problem.

Advanced password cracking techniques like hybrid attack and combinator attack further undermine the security of user-chosen password scheme [26]. Hybrid attack combines dictionary and brute-force attacks to perform a more efficient cracking, whereas combinator attack combines every word in a dictionary with every other word in a dictionary to guess users' passwords. Passwords of reasonable length with a combination of uppercase letters, digits, and special characters (e.g., "Letmein1!") could easily be cracked by using these sophisticated techniques.

2.2 Password Reuse and Password Similarity

The phrase "domino effect of password reuse" was first coined in the work of Ives et al., who speculated that a domino effect might occur as a result of one site's password file falling prey to a hacker [57]. They conjectured that the hacker might exploit that file to try infiltrate other systems as well, and the habit of password reuse across different sites would certainly make the job of the hacker easier. They predicted that with the proliferation of password-protected accounts, users would reuse passwords more across different sites and the scenario might get worse.

Notoatmodjo and Thomborson surveyed a group of users and found out that they mentally group their accounts [83]. They identify the factors based upon which these groupings are made and also show that password reuse rate is greater for accounts that are considered less important than accounts that are considered more important.

Their classification of user accounts in terms of perceived importance level is vague, since it consists of only two groups—“less important” and “more important” accounts, and they only focus on password reuse without any kind of modification. We proposed a more concrete classification of user accounts and examined other significant issues, such as reuse with some modifications and reuse using a similar thought process. Notoatmodjo and Thomborson also argue that password reuse is a good strategy for less important accounts, reserving mental capacity for more important accounts. While we agree with this notion, we found that users exhibit both partial and complete password reuse between less and more important accounts, creating a serious increased risk for the user.

Preibusch and Bonneau used a game theoretic model to explain the password schemes used by security-indifferent and security-concerned websites [93]. In another work, they performed a large-scale comparative analysis of password implementation strategies of websites of different categories [12]. Komanduri et al. conducted a user study to examine how different password-composition policies for different websites actually affect the users [68].

The only prior work we found on password similarity was that of Zhang et al., who studied password expiration and the relationship between users’ previous passwords and their new passwords [128]. They examined over 7700 accounts of a Single Sign-On system of a university and their results demonstrated that old passwords are effective predictors of new passwords. We examined similarity of passwords from different accounts of different classes.

Kaye investigated password sharing practices of users through a self-report measure, in which one-third of the participants reported that they shared their personal email password, while a quarter reported that they shared their Facebook password, mainly with partners and close friends [64]. A part of our work was devoted to in-

investigating shared passwords, but we mainly explored the hidden consequences of password sharing by investigating the extent to which shared passwords are reused elsewhere. For example, when users share their Netflix passwords with friends, they are consciously doing it, but if they reuse the same Netflix password for their personal email accounts, they inadvertently create a potential breach in the privacy of their personal email accounts. Kaye investigated the former case while we focused on the latter.

Chiasson et al. conducted a laboratory study to compare the recall success rate of textual passwords with that of graphical passwords [21]. Their experimental methodology involved constructing textual passwords for six different kinds of accounts: bank, email, instant messenger, library, dating, and work. These passwords conceptually belong to the higher level of our hierarchy. They performed a visual inspection of these higher-level passwords and observed that most of the participants constructed passwords following a common pattern across their accounts. Their results also demonstrated that, compared to graphical passwords, recall success rate was lower for textual passwords. By conducting all these tests, they highlighted the weaknesses of textual passwords and advocated the effectiveness of graphical passwords.

Although our experimental methodology has some similarities with Chiasson et al.'s study, their experimental hypotheses were completely different. They focused on comparing between textual and graphical passwords, whereas we focused on comparing between textual passwords of different importance levels. The spectrum of constructed passwords in their experiment did not contain any lower-level passwords. On the other hand, we asked our participants to create passwords of different importance levels and exploited the lower-level passwords of a specific user to crack that user's higher-level passwords.

Adams et al. conducted a web-based survey with 139 participants to investigate the usability issues in password systems [2]. They find out that the memorability of a particular password is significantly correlated with its frequency of use. Memorability is also significantly correlated with automaticity or the ability to recall a password spontaneously without conscious thinking. They argue that these findings are consistent with cognitive theory principles like *encoding specificity* and *explicit vs. implicit* memory models. They further conducted semi-structured interview sessions with 30 participants to examine a few important issues more deeply and use grounded theory from social sciences to analyze the responses and build a model of users' password behavior.

In contrast, we first proposed a model of user password hierarchy based on certain observations and then verify the model by collecting data from users. We also analyzed the similarity between the lower and higher level passwords in our hierarchy. The preliminary evidence of such similarity is indicated by another important finding of their study which highlights that almost half of the users have a common theme for all or most of their passwords [2].

Many other novel methods have been used for understanding user password habits and attitudes. Hayashi and Hong used a diary study [54], Florêncio and Herley installed an opt-in component of the Windows Live Toolbar in users' machines [36], Shay et al. capitalized on the opportunity of a Carnegie Mellon University (CMU) password policy change [108], while Gaw and Felten gathered feedbacks from users after they had made actual login attempts in different websites [41].

Our work differs from all these works in two major ways. First, we proposed a concrete categorization of password-protected sites and presented our hypotheses and research questions based on this categorization. We tested these hypotheses by both collecting passwords from users and reviewing their responses to a questionnaire.

Second, we did not observe the degree of reuse only, rather we observed the degree of similarity among passwords used at different levels of our proposed hierarchy. Therefore, when designing our survey, we considered all possible similarities (both syntactic and semantic) among multiple passwords of a user and prepared our questionnaire accordingly.

2.3 Mobile Device Password Entry

Prior research has shown that text entry requires more effort on mobile phones. Bao et al. report that typing speed is significantly slower on phones than on PCs [8]. General observation suggests that capitalizing a letter and inserting digits/special characters also require more effort on a phone than on a computer. To date, however, no empirical work has examined how these factors actually affect the construction of passwords on computers and mobile phones. To the best of our knowledge, ours was the first empirical study that examines how password strengths vary across computer keyboards and mobile phones of different layouts.

Our proposed custom layout also provides a novel mechanism for inserting digits and special characters when constructing a password. It removes the burden of making an extra click when inserting digits/special characters in a password. Related works have mainly focused on improving the general typing speed on mobile phones. For example, proposals have been made for adding *chording* to numeric feature phone keypads [91, 122]. Other works have focused on pressure-based text entry [15, 113], but pressure-based schemes are often error-prone. Chiang et al. and Schaub et al. evaluated the usability of graphical password schemes on smartphones [20, 101], but graphical passwords have yet to replace textual passwords as the primary authentication mechanism in most systems.

In their work, Jakobsson and Akavipat took advantage of the auto-correction and auto-completion features of mobile handsets and implemented a mechanism called *fastword* [59], which is two to three times faster to enter than an ordinary password. Their experimental results showed that fastwords have greater entropy and higher recall rates than ordinary passwords. *Fastword*, unfortunately, is not fully compatible with existing sites that have arbitrary limits on password length and may require special characters, digits, and capital letters. Our designed custom layout aimed to assist users to enter digits and special characters in a more convenient way.

2.4 Questionnaire Development

2.4.1 Usability Questionnaires in HCI

In HCI, efforts to develop usability questionnaires have been limited to the domain of software product evaluation only. Early researchers developed questionnaires such as SUMI (Software Usability Measurement Inventory) and QUIS (Questionnaire for User Interaction Satisfaction) for measuring software quality from the end user's point of view [67, 24]. Later on, these questionnaires were pointed out to be too generic [70]. As a result, researchers started to develop more specific questionnaires tailored to particular groups of software products. Examples of this kind of questionnaire are WAMI (Website Analysis and Measurement Inventory), MUMMS (Measuring Usability of Multi-Media Systems), and UFOS (Usability Questionnaire for Online Shops) [66, 70].

Since all these questionnaires focus on software products, they are not helpful for evaluating the password interface of a mobile device. The key component of a password interface is the keypad layout through which the password is typed. Layout issues like capitalizing method, inter-key distance etc. are important considerations

here, which can not be captured by existing software product evaluation questionnaires.

Due to the proliferation of mobile devices and the requirement of a usability questionnaire that is specific to this technology, a questionnaire has been particularly developed for evaluating mobile user interface. Ryu and Smith-Jackson developed a usability questionnaire for mobile devices which was subsequently shortened (Mobile Phone Usability Questionnaire, also known as MPUQ) to improve reliability and validity [99]. However, their questionnaire does not contain any item that addresses the issue of mobile password entry. Although typical tasks for mobile phones such as checking missed calls, sending/receiving short messages etc. were considered in the questionnaire, password entry as well as other security tasks were overlooked.

2.4.2 Psychometrics in HCI

In the existing literature of usable security, we have not found any instance of applying psychometrics to solve a particular usability problem. However, HCI researchers have adopted psychometric theory approaches to measure user satisfaction. The above mentioned questionnaires such as SUMI, QUIS, and MPUQ were developed by following psychometric theory approaches.

In his work, Lewis evaluates the psychometric properties of four existing IBM questionnaires that were developed for measuring user satisfaction with computer system usability [74]. He provided the questionnaires to different users after they had completed certain computer tasks and asked them to express their opinion about the computer system they had just interacted with. By analyzing the response scores and measuring the reliability and validity, he concluded that all the questionnaires have acceptable psychometric properties, thus allowing the usability practitioners to use them with confidence for measuring user satisfaction with different computer systems.

2.4.3 Psychological Approach in Usable Security

Our effort of applying psychometrics in solving a usability problem is inspired from the observation that psychological approaches have been useful to solve usable security problems. A notable example of this is the work of Jaferian et al., in which they apply *activity theory*, a revolutionary theory originating in Soviet psychology [73], to develop a set of heuristics for evaluating the usability of IT security management tools [58]. Their results demonstrated that the heuristics performed well in identifying usability problems in IT security management tools.

2.5 System-assigned Random Passwords

Multiple study results reported the poor memorability problem for system-assigned random password scheme, even when natural-language words are used [106, 123]. In their study, Wright et al. compared the memorability of three different system-assigned random password schemes: Letter Recall, Word Recall, and Word Recognition [123]. The memorability results were not encouraging for any of the schemes, which confirm the poor memorability issue of system-assigned random passwords.

Researchers have also tried to find a compromise between user-chosen and system-assigned passwords. Forget et al. proposed the Persuasive Text Passwords (PTP) scheme [39], in which users are first asked to choose their own passwords. Next, the password is modified by the system by inserting random characters at random positions. However, they did not conduct a multi-session study to assess the memorability of these passwords [39].

CHAPTER 3

Hierarchy of Users' Web Passwords: Perceptions, Practices, and Susceptibilities

As we are interested in examining the degree of similarity among passwords of different importance levels of an individual user, we first propose a user password hierarchy based on the perceived importance of the site. In the context of this work, the term “importance” denotes how much effort a user provides to protect the security and privacy of a password. This includes activities such as constructing a strong password that would be hard to guess, not writing the password on a piece of paper, not sharing it with others etc. Since users do more to protect the passwords that they consider more important, this hierarchy is also a password hierarchy based on the level of privacy that the users give their passwords.

First, we classify all the password-protected accounts of a user into five broad categories. Whereas the first three of them are adopted from the categorization of Bonneau and Preibusch [12], we further add two more categories. We now present a brief description of each category.

Identity Accounts In today’s digitalized world, the virtual identity of a user has become more and more important. A user’s official webmail account acts as the medium of her professional correspondence, her social networking account personifies herself among friends and family members, and her blogging account represents her voice about different issues. A user creates online identities in these sites that act on her behalf. She builds a long-term reputation of trust in her professional and personal

life through them. In short, a user has significant incentive to protect the security of these accounts.

Financial Accounts For online banking and investing, users need to create online accounts for various kinds of financial transactions and bill payments. Users are always concerned about the security of these financial accounts because they represent the users' access to their money and credit. Compromise of these accounts may reveal credit card information and other financial credentials. We consider online banking accounts and accounts in all kinds of merchant sites as financial accounts. Users try to ensure maximum protection for these financial accounts.

Content Accounts Users create accounts in some websites only to customize the contents of those sites. In these accounts, users do not have significant interactions with other users nor any financial transactions. For example, if a user wants Weather.com to show the weather of her location when she visits the site, she might need to create an account there. News and other informational websites also belong to this category. If an account of this kind is hacked, it does not pose much threat to the user.

Sketchy Accounts It is unlikely that all of a user's password-protected accounts belong to a category of well-recognized identity, financial, or content sites as described above. In our study, we consider users' accounts in any kind of unrecognized websites as sketchy accounts. This category includes unfamiliar sites that claim to have various kinds of deals or coupons and little-known online forums or content provider sites. Users create accounts in these sites for superficial purposes and often they maintain anonymity by providing a false name or age. As a result, compromise of any account

of this kind usually does not create any breach to their privacy. Users have the least incentive to protect the security of these accounts.

Shared Accounts Sharing accounts among multiple users is a common practice. In most cases, apartment mates share the same wireless Internet account. Accounts are also shared to a large extent on paid subscription websites that offer paid access to premium content. Users who want to save money split the subscription fees of these sites and share their member accounts. The password associated with a shared account is known to all of the members who share the account. Passwords are shared for both identity (e.g., colleagues may share the password of a work email account) and financial accounts (e.g., spouses may share the password of a bank account) [64].

3.1 Hierarchy

If we consider the privacy of passwords for these accounts, it appears that users maintain maximum privacy for their financial and identity passwords. There exists no empirical work in the current literature that shows which of these two types of account is more important to users. A financial account (online banking account, for example) is certainly of great importance, but having access to an identity account (email account, for example) also often means getting access to other accounts that are linked to that identity account (by sending a password reset request mail to the linked email account). We therefore consider financial and identity passwords to be equally important. If these passwords are leaked, serious consequences may result.

On the other hand, passwords for shared accounts are constructed for the purpose of sharing them with others. Generally, passwords are shared among close friends or family members, and people who share passwords also share a level of trust. We predict that users do not create these passwords with the thought that they would

share them outside their close circle friends or family members in the future. They create these passwords according to their own criteria, and then sometimes share them due to expedience or circumstance. For example, a user shares a wireless Internet password with her trusted long-term next-door neighbor. Suddenly, the neighbor moves on and a new neighbor comes to the apartment and continues sharing the existing Internet connection. The new neighbor is just a casual acquaintance whom the user barely knows. Now the new neighbor already knows the shared wireless Internet password and other credentials of the user (e.g., the email ID that the user uses to forward the bill payment receipt to the neighbor). If the neighbor has malicious intent, she may try to hack the email account of the user by using the email ID and the wireless password.

Similarly, passwords for sketchy accounts can also be exploited for compromising other important accounts. A sketchy account is created on an unrecognized website, which may be purposefully designed as part of a social engineering scheme. This scheme exploits the fact that users tend to create accounts on new websites somewhat indiscriminately. An attacker could thus create a website that provides a simple web service and recommend that a new user should create an account on the site to gain full access to that service. In this way, the attacker collects users' passwords and other credentials like usernames and email addresses. Subsequently, the attacker tries to hack accounts on other common financial or identity sites by using this information. If a user reuses the same combination on those important websites, her accounts on those websites would be compromised by the attacker.

In the light of the above discussion, it can be seen that financial and identity passwords are the target passwords that an attacker would like to crack, while shared and sketchy passwords are the passwords that an attacker might want to exploit. Unlike shared or sketchy passwords, content passwords are not readily available to

a potential attacker. As mentioned before, content passwords are created in well-recognized trusted websites (New York Times, for example). However, compared to financial and identity websites, content websites do not require a high password security level because these sites do not protect sensitive personal or financial information for their users. As opposed to a financial or identity password, the potential harm that is caused by the leakage of a content password is nominal. Thus, neither the user, nor the site authority, has much incentive to protect the privacy and security of a content password.

We therefore propose a user password hierarchy by placing financial and identity passwords at the higher level, and content and sketchy passwords at the lower level. Since shared passwords can belong to multiple categories, we do not include them in our hierarchy. However, in our user study, we add questions about shared passwords in order to explore the hidden consequences of password sharing. Figure 3.1 illustrates this hierarchy.

3.2 Hypotheses and Research Questions

One important objective of our study is to test the validity of our proposed hierarchy regarding financial, identity, content and sketchy passwords.

Hypothesis 1 *In terms of the perceived importance of the sites, a hierarchy of users' Web passwords can be proposed, where financial and identity passwords sit at the top level, while content and sketchy passwords sit at the bottom level of the hierarchy.*

The next and the most important objective of our study is to examine how the knowledge of a password of a lower-level account (content or sketchy account) could increase the chance to crack a higher-level account (identity or financial account) based on similarity to the lower-level password.

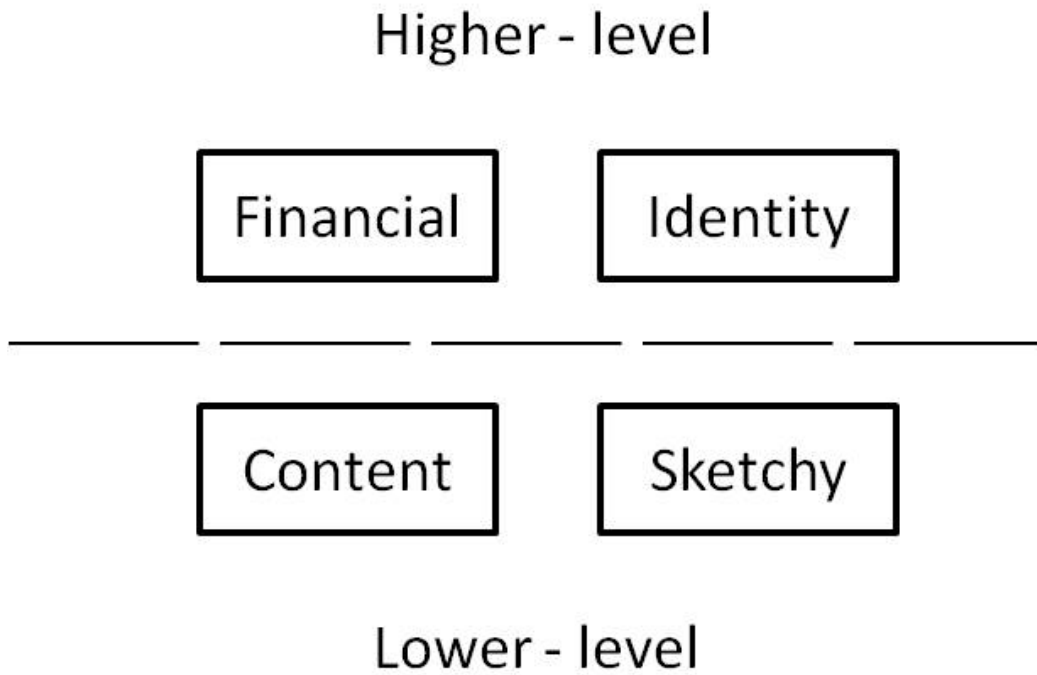


Figure 3.1: User password hierarchy based on the perceived importance level.

Due to the prevalence of content and sketchy sites, users frequently encounter these sites while surfing the Internet. Therefore, we hypothesize that most users maintain a fixed set of passwords for these unimportant sites so that they do not need to create and remember a new password each time they create a new account on these sites. These users are perceived to be more careful and we hypothesize that they usually do not reuse this fixed set of passwords in their financial or identity accounts.

Hypothesis 2.a *Most of the users a) use a fixed set of passwords for sketchy sites, and b) of those that do, they usually do not reuse this fixed set of passwords in their important financial or identity accounts.*

Hypothesis 2.b *Most of the users a) use a fixed set of passwords for content sites, and b) of those that do, they usually do not reuse this fixed set of passwords in their important financial or identity accounts.*

On the other hand, users who do not maintain a fixed set of passwords for content or sketchy sites need to create new passwords frequently. As discussed before, the cognitive capacity of a typical user restricts the user from constructing a new random password from scratch every time because it is not possible for the user to remember so many random passwords. We therefore predict that when creating the new password, users either reuse one of their existing passwords (with or without modifications), or they use a similar process as they have used before to create one of their existing passwords.

However, what remains quite unclear is to what extent they reuse their important financial or identity passwords (with or without modifications), or to what extent they use a similar process as they have used before to create one of their existing financial or identity passwords. We address this as an open research question that we try to answer through our user study.

Research Question 1 *When creating a new password for a sketchy account or a content account, a) to what extent users reuse one of their financial or identity passwords, without any modification, b) to what extent users reuse one of their financial or identity passwords, with some modifications, and c) to what extent users use a similar process as they have used before to create one of their existing financial or identity passwords?*

Another objective of our study is to explore the degree of similarity between shared passwords and higher-level passwords. We seek to learn the extent to which users reuse their higher-level (financial or identity) passwords for creating a shared password. This leads to the formulation of our second research question.

Research Question 2 *When creating a password for a shared account, a) to what extent users reuse one of their financial or identity passwords, without any modification, b) to what extent users reuse one of their financial or identity passwords, with some modifications, and c) to what extent users use a similar process as they have used before to create one of their existing financial or identity passwords?*

3.3 Methodology

We conducted a computer-based two-phase laboratory study with 80 UTA students to test our research hypotheses and answer our research questions. In the first phase of the study, we asked the participants to construct new passwords for websites of different categories. This phase was hosted on the secure web server run by the Information Security (iSec) Lab at UTA. Once this phase was completed, each participant was redirected to www.surveymonkey.com¹ for the second phase. In this phase, we had the participants answer some questions regarding their password behaviors for multiple accounts.

Although a larger number of participants could have been drawn from an online survey, we preferred a laboratory study because our pilot study (n=12) showed that a laboratory study would produce more consistent responses, especially in the first phase, where the students would be asked to create passwords for eight different websites. Students were assigned partial course credit in exchange for their participation. The complete study was approved by the UTA Institutional Review Board (IRB). With prior approval from UTA IRB, electronic informed consent was obtained from the participants in lieu of written informed consent. After analyzing the passwords

¹SurveyMonkey is a website for administering and analyzing online surveys.

constructed by the participants, we encrypted and stored them in a disk disconnected from any kind of network.

3.3.1 Study Administration

We administered the study through the research pool of the department of psychology, UTA. The department of psychology at UTA maintains the pool for assigning partial course credits to the students who enroll for the course “Introduction of Psychology” and for some other advanced elective courses that offer extra credits. Researchers who collaborate with the department of psychology can post a brief description about their studies to the pool. Students in the research pool can view all the studies and sign up for those that interest them.

The main advantage of conducting a study through the pool is that it can draw a wide range of participants from various departments, because most of these courses are offered for majors from all departments. However, before the beginning of the study, we explicitly informed each participant that the study was being conducted by the Information Security Lab. This was done so that no participant would confuse our study as an experiment for measuring the psychological aspects of people through their constructed passwords.

3.3.2 First Phase

The main objective of the first phase of the experiment was to capture multiple passwords of a user so that we could examine the degree of similarity among them. For this purpose, we designed a PHP script that prompted the users to create passwords for their new accounts for eight different websites in four different categories:

- Financial website: Chase and Wells Fargo
- Identity website: Yahoo! Mail and Facebook

- Content website: NY Times and Weather.com
- Sketchy website: Dreamdeals.com and Justchill.com (hypothetically constructed sites)

We selected Chase and Wells Fargo as representatives of banking/financial websites because these two banks should be familiar to UTA students due to the prevalence of their ATMs on the campus. Facebook and Yahoo! Mail were selected as identity websites, mainly because of their popularity as a social networking site and a webmail site, respectively. For content websites, we selected the NY Times website and Weather.com, because these two sites readily present a clear distinction between identity sites and content sites, without us needing to explicitly label them as content sites.

During the first phase, we did not want to give the participants any clue about our experimental motive because we expected them to spontaneously construct new passwords, exactly in the same way as they do in real life. Therefore, for all the six real sites, we designed the interfaces so that they would look similar to the original sites. For the two hypothetical unfamiliar sketchy sites, we gave their interfaces an informal appearance so that they would appear to the participants as real-world sketchy websites.

3.3.2.1 Password Construction

In our study, we did not read the participants a script or provide them any written instruction. The instructions were presented on the computer screen. For ethical and security reasons, we explicitly told the participants through warnings in our interface not to provide any of their existing passwords. For each website, we provided a brief introduction and presented a real-life scenario to the participants. The scenario was created in such a way that it resembled a real-world application

as much as possible. For example, for Weather.com, the participants were presented with the following scenario:

Weather.com provides the latest weather forecasts, maps, and alerts. You want Weather.com to show weather for Arlington, TX when you go to the site. To do that, you need to register an account on Weather.com so that you can customize your location. Imagine that you are registering a new account on Weather.com. You have reached the final step of registering your new account, and you need to input a password. Proceed to the next page to input your new password.

Once the user clicked the link, our mock password construction page for Weather.com appeared. Figure 3.2 shows the interface for the mock password construction page for Weather.com. We note that the URL of the webpages (the URL of our secure web server that hosted the code) appeared in the browser and the participants were aware that it was not the actual Weather.com password construction page. We also anticipated the fact that some participants might consider Weather.com as a sketchy site (due to their unfamiliarity with Weather.com). We believe that our short description of the sites helped to remove this kind of misconceptions from the participants.

Similarly, for Dreamdeals.com, the participants were presented with the following scenario:

Imagine that you are doing a Google search to find discount coupons for the Six Flags amusement park in Arlington. While browsing multiple search results pages, you come across a site called Dreamdeals.com that offers coupon codes for Six Flags and also deals, discounts, and cash backs for other purchases. Dreamdeals.com requires you to register an account in order to gain access to the coupons and discount codes. Imagine that you are registering an account on Dreamdeals.com. You have reached the

Profile Registration

Complete your registration

Already registered? Sign in!
Forgot password?

Please construct the password for your new profile and complete your registration

TRUSTe
CERTIFIED PRIVACY

Easy Access
Get easy access to other
registration products
Save Your Profile
Become a member of
the weather.com club

Password*

Confirm Password*

Save

* required fields

secure site

By submitting your information, you indicate that you have read, understand, and agree to the weather.com Legal Restrictions and Terms of Use and Privacy Policy, and that you consent to the collection and use of this information and the transfer of this information for processing and storage by weather.com and its affiliates.

Copyright © 1995-2011, The Weather Channel Interactive, Inc. Your use of this site constitutes your acceptance of the LEGAL RESTRICTIONS AND TERMS OF USE

UPDATED weather.com © Privacy Statement - Licensed by TRUSTe | Parental TV Controls

Figure 3.2: Mock password construction page for Weather.com.

final step of registration, and you need to input a password. Proceed to the next page to input your new password.

As mentioned before, we created an informal looking interface for Dreamdeals.com, as shown in Figure 3.3.

3.3.2.2 Password Policy

For all six real websites, we enforced exactly the same password policies as they are enforced in those sites. For example, Wellsfargo requires any password to be 6-14 characters long, with at least one letter and one digit. We designed our script in such a way so that the participants had to conform to this policy. For the two hypothetical sites, we ensured that the participants' passwords were at least five characters long.

dreamdeals.com

Search for Coupons, Deals, Stores, and Offers **Search**

Coupons & Deals Forums **Cash Back**

Complete your registration Already registered? Sign in! Forgot password?

Please construct the password for your new profile and complete your registration

Password*

Confirm Password*

* required fields **Save**

DEAL

ELECTRONICS **HEALTH AND BEAUTY** **KITCHEN APPLIANCES** **KIDS AND FAMILY**

About Us | Blog | Site Map | Mobile | Contact Us | Partnership | Privacy | User Agreement | D.M.C.A. Notice | Civil Process Policy

Figure 3.3: Mock password construction page for Dreamdeals.com.

Like the original sites, participants were also required to reconfirm their passwords in a second box, which prevented them from typing some random characters as their passwords.

In this way, we implicitly tried to trigger the real life password creation mechanisms of users for websites of different categories. In designing the interfaces and providing the introduction for each site, we were careful about not revealing to the participants that our main objective is to categorize their constructed passwords based on our categories. We believe that this helped to avoid any kind of experimental bias that is associated with “demand characteristics” [87].

3.3.3 Second Phase

In the second phase of our experiment, we asked the participants to answer a survey. In this phase, we were relatively overt about our categorization of passwords. We asked users to contemplatively respond to some questions about their password sharing habits and password reuse habits, with and without modifications, across websites of different categories.

We were aware of the fact that it would not be a straightforward task for the participants to exactly recall the construction processes of passwords they had set up some time ago. Moreover, users are not actively aware of the processes carried out during password construction since password construction is not the primary task of users, rather it is one of a series of subtasks required for completing the primary task (opening a bank account, for example). For this reason, instead of asking them to provide open-ended responses, we gave them a series of options and asked them to select those options that are related to their password construction processes. The options were carefully selected from prior studies. We believe that these options provided cues for the participants to recall the processes they usually carry out during their password construction activity. We give a brief overview of our survey questions here.

3.3.3.1 Rating of Sites

We asked the participants to rate the importance of their passwords on a 5-point Likert-type scale for all eight websites that were presented in the first phase. We were aware of the argument on whether responses from Likert scales should be considered as ordered-categorical data or interval-level data [27]. To treat the scale as an interval-level scale, anchors were only included on the bipolar ends of the scale (1=“not important”, 5=“important”), and the middle point (3=“moderately important”).

3.3.3.2 Shared Passwords

Next, we gave the participants two real-word examples of password sharing: one about sharing Netflix password with friends and the other about sharing a wireless Internet password with apartment mates. We asked them whether or not they share their passwords for cases like these. Those who responded that they do share were then asked how they create passwords for these shared accounts. We first asked them to what extent they reuse a password they have used elsewhere, without any modification. We used a 4-point Likert scale (1=“never”, 2=“seldom”, 3=“sometimes”, 4=“often”) for this and the subsequent questions. Those who responded “sometimes” or “often” were further asked which kind of password they reuse.

We then asked them the extent to which, when creating a shared password, they reuse a password they have used elsewhere, with some modifications. More specifically, we asked them “When creating a password for your shared account, how frequently do you reuse a password that you have used elsewhere, with some modifications (for example, $abc \rightarrow abc1$, $abc \rightarrow abd$ etc.)?” The examples were intended to limit the different ways that different users might interpret “modifications” to mainly focus on minor changes, such as adding or changing just a single character. Those who responded that they reuse a password with some modifications were further asked which kind of password they reuse with modifications, and what kind of modifications they make. Finally, we asked them to what extent they use a similar process as they have used before when creating shared passwords. Those who responded that they use a similar process were further asked which kind of password would be most similar to the new shared password, and how they would be similar.

3.3.3.3 Sketchy Passwords

In this part, we first revealed to the participants that the two sites Dreamdeals.com and Justchill.com that were presented before belonged to our category of unfamiliar sketchy sites. Then we asked the participants whether, for this kind of site, they use a fixed set of passwords or create a new password each time they open a new account for each site.

The participants who responded that they use a fixed set of passwords were further asked whether they reuse this fixed set of passwords elsewhere, especially in their identity/financial websites. On the other hand, those who responded that they create a new password each time were further asked the same questions as they were asked about shared passwords. They were asked the extent to which, when creating a new sketchy password, they reuse a password they have used elsewhere, use a similar process as they have used before, and so on.

3.3.3.4 Content Passwords

As with sketchy passwords, we first revealed to the participants that the two sites NY Times and Weather.com that were presented before belonged to our category of familiar content sites. Then they were asked the same questions about their content passwords as they were asked for sketchy passwords.

3.3.4 Survey Analysis

To test our hypotheses, we analyzed the responses of the participants from both phases. We first thoroughly reviewed their responses to the questionnaire that was provided in the second phase of the study. Then we collected the passwords that the participants constructed in the first phase and divided them by category. We analyzed

each group separately to find out the frequency of using capital letters, digits and special characters.

Finally, we tried cracking the higher-level (financial and identity) passwords with the help of lower-level (content and sketchy) passwords. We used the John The Ripper (JTR) password cracker for this purpose, using JTR's *wordlist* mode combined with the *single crack* mode.

Wordlist mode cracking is basically a dictionary attack where every word in a wordlist is tried against the candidate password until a match is found. If word mangling rules are enabled, each word in the wordlist is modified or mangled to generate other possible combinations. The single crack mode is the default cracking mode of JTR in which a large number of word mangling rules are applied to a very small dictionary to perform a dictionary attack. As the default set of word mangling rules is very small in the wordlist mode, we modified the JTR configuration file so that it would be possible to apply the large set of word mangling rules of the single crack mode while performing cracking in the wordlist mode.

For each participant, we combined the participant's lower-level passwords with a comprehensive wordlist and tried to crack the higher-level passwords of the same participant by using JTR in our modified wordlist mode.

3.4 Results

In this section we present all the major findings based on our analysis of 80 surveys.

3.4.1 Demographics

Overall, 47 female and 33 male students participated in our survey. Students from diverse majors, including Psychology (13), Nursing (12), Kinesiology (6), Biology

(5), Engineering (5), Business (4), Education (4), Social Work (4), and others (27) participated in our survey.

3.4.2 Perceived Importance of Passwords

Our first hypothesis was that users mentally classify all of their passwords into different categories according to their perceived importance. By analyzing the ratings provided by the participants (Section 3.3.3.1), we found evidence to support this hypothesis. The identity and financial passwords were perceived to be more important (had higher mean and median ratings) than the content and sketchy passwords. Figure 3.4 summarizes the ratings of the participants for all eight sites.

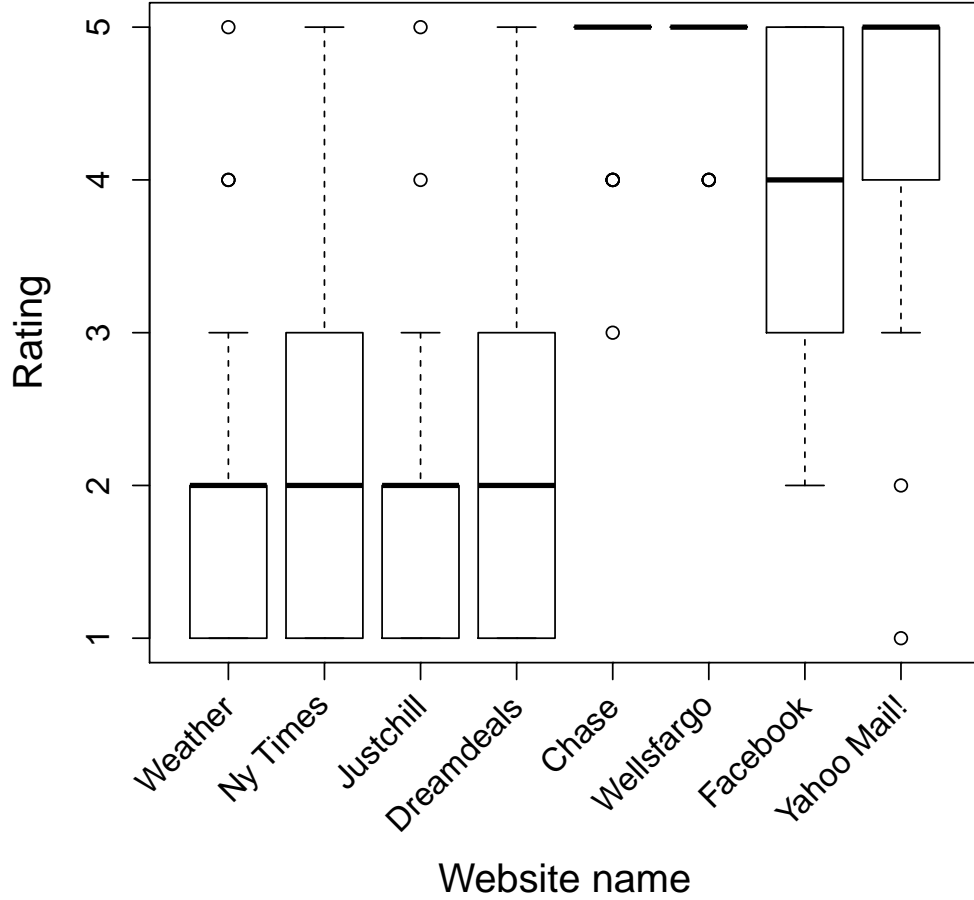


Figure 3.4: Box plot of participant ratings on a 1 to 5 scale (1 being “not important”, 3 being “moderately important”, 5 being “important”) about the perceived importance of passwords of different websites.

Given the skewness evident in Figure 3.4, we decided to conduct a nonparametric repeated measure statistical test to properly examine Hypothesis 1. Accordingly, we conducted Friedman’s test. The perceived importance of passwords differed significantly across the eight sites, $\chi^2(7) = 448.017, p < .001$.

Table 3.1: Summary of post hoc analysis with Wilcoxon signed-rank tests with a Bonferroni correction applied. For brevity, all entries with $p < .00179$ have been omitted.

Pair	Z	Asymp. Sig. (2-tailed)
Wellsfargo-Chase	-1.342	.180
Dreamdeals.com-Justchill.com	-2.238	.025
Dreamdeals.com-NY Times	-.256	.798
Dreamdeals.com-Weather.com	-2.753	.006
Yahoo! Mail-Facebook	-2.962	.003
Justchill.com-NY Times	-2.207	.027
Justchill.com-Weather.com	-1.207	.228

Post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance level set at $p < 0.0179$. All the banking and identity sites had significantly higher ratings than all the content and sketchy sites, $p < .00179$ for all cases. The banking sites also had significantly higher ratings than the identity sites, $p < .00179$ for all cases. The content site Weather.com had a marginally significant lower rating from the content site NY Times, $p < .00179$. The differences between other pairs of content and sketchy sites were statistically insignificant. The difference between the webmail site (Yahoo Mail!) and the social networking site (Facebook) was also statistically insignificant. Table 3.1 summarizes the results for the post hoc analysis.

The above results suggest that, although financial and identity passwords sit at the top level of the password hierarchy according to their perceived importance, financial passwords are perceived to be significantly more important than identity passwords. On the other hand, content and sketchy passwords sit at the bottom level of the hierarchy. However, there was not enough evidence to make any clear distinction between content sites and sketchy sites.

The hierarchy of users' web passwords, therefore, turned out to be a three-level one (as opposed to a two-level hierarchy as we hypothesized), where financial passwords sit at the top level and identity passwords sit at the next level, while content and sketchy passwords sit at the bottom level of the hierarchy.

3.4.2.1 Password Characteristics

We now analyze the passwords constructed by the participants in the first phase to help validate the findings from the survey in the second phase. Our findings are also consistent with Hypothesis 1.

We calculated the length of the passwords and the frequency of using capital letters, digits, and special characters for passwords of different categories. The length and the frequency values decreased as the perceived importance of the sites decreased. Figures 3.5 and 3.6 summarize our analysis.

We begin with an analysis of password length, shown in Figure 3.5. Passwords are longer for financial sites and then shorter for identity, content, and sketchy sites, in order. We note that the minimum password length requirement was not the same for all the sites (seven for Chase, five for NY Times, Justchill.com, and Dreamdeals.com, six for all others), and this may have affected the password lengths.

The frequency of using capital letters also decreased from higher-level passwords to lower-level passwords (Figure 3.6). Unlike length, no confounding effect existed in this case because the participants were not required to use capital letters in any of the sites. They spontaneously used more capital letters when constructing their higher-level (financial and identity) passwords.

Among the eight sites, only the two financial sites required their passwords to contain at least one digit. Yet the frequency of using digits decreased from identity sites to sketchy sites (Figure 3.6). The same is true regarding the use of special char-

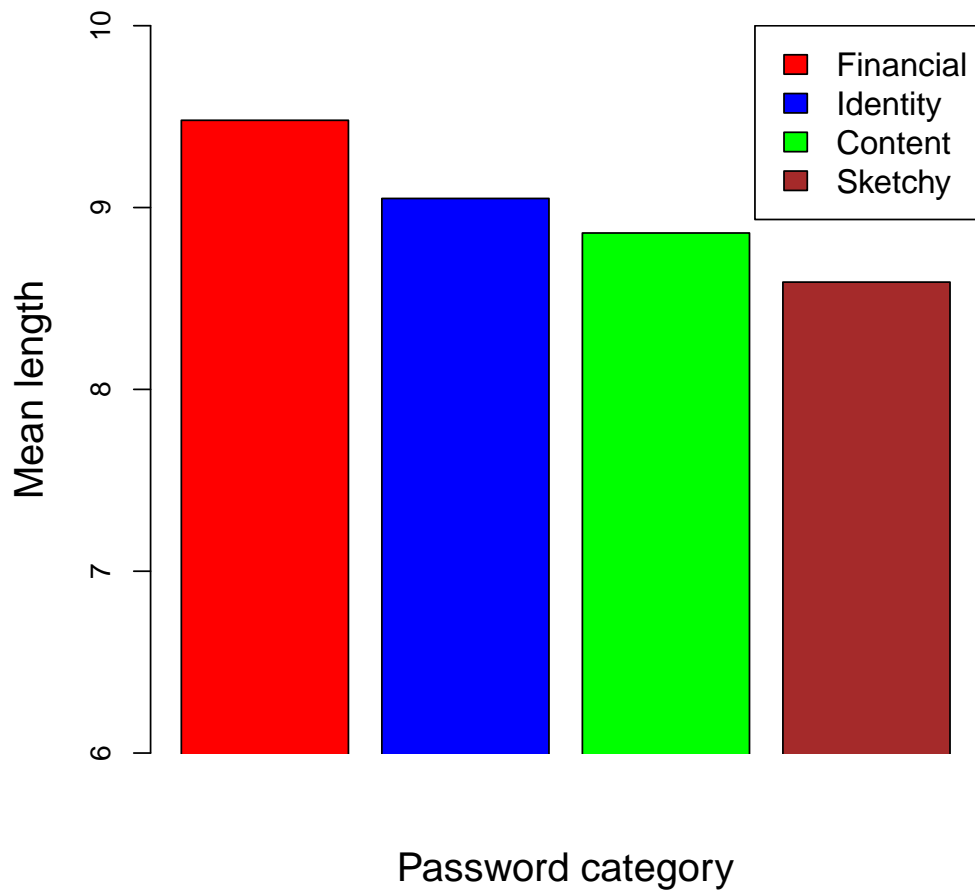


Figure 3.5: A comparison of mean lengths.

acters. None of the identity, content, or sketchy sites had any kind of requirement or restriction for using special characters. Still the frequency of using special characters decreased from identity sites to sketchy sites (Figure 3.6). Since Chase does not allow special characters for their website and we followed the password policy of the actual websites, we did not allow any special characters to be included in the passwords constructed for Chase. Therefore, the frequency of using special characters was less in the financial sites than the identity sites (Figure 3.6).

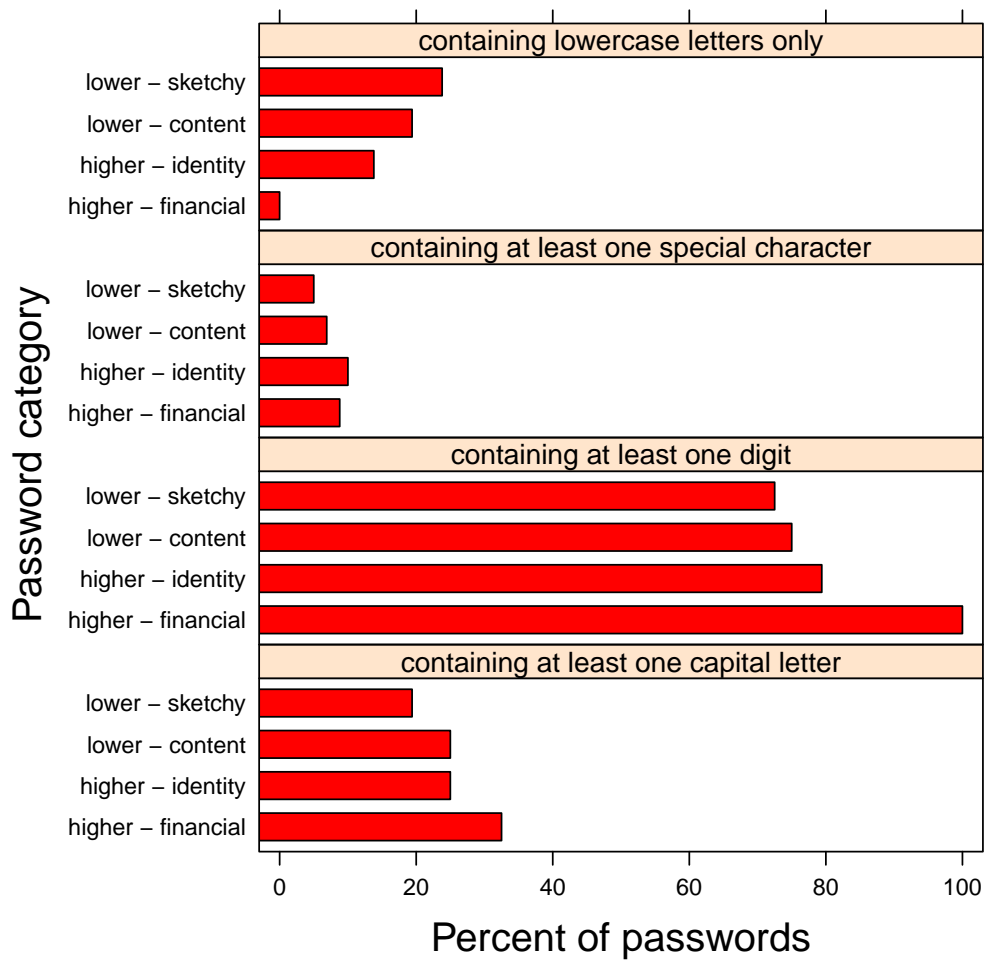


Figure 3.6: A comparison among passwords of different categories.

We also calculated the percentage of passwords that contained only lowercase letters, with no capital letters, digits, or special characters. The percentage increased from identity sites to sketchy sites (Figure 3.6), although the password policy was same for all the identity, content, and sketchy sites regarding the use of capital letters, digits and special characters.

3.4.3 Sketchy Password

We report our results about sketchy passwords by relating them to relevant hypothesis/research question.

Hypothesis 2.a

As suggested in the first part of Hypothesis 2.a, it was found that most of the users use a fixed set of passwords for sketchy sites. Specifically, 52 of our 80 participants (65%) reported using a fixed set of passwords for sketchy sites whereas 28 reported that they create a new password each time they open a new account at an unfamiliar sketchy site. However, the chi-square statistics demonstrated that the difference was not significant, $\chi^2(1) = 3.09$, $p = .079$.

For the second part of the Hypothesis 2.a, it was proposed that of the participants that do use this fixed set of passwords that they usually do not reuse this fixed set of passwords in their important financial or identity accounts. To assess this, we compared participants who reported “never” to “seldom” reusing against those that “sometimes” to “often” reused. This part of the hypothesis was supported, 37 of these 52 participants reported “never” to “seldom”, whereas only 15 participants reported “sometimes” to “often” reusing these fixed passwords for their financial or identity accounts, $\chi^2(1) = 4.03$, $p < .05$.

Research Question 1

Overall, 28 of our 80 participants reported that they create a new password for a sketchy site. As mentioned in Section 3.2, it was not quite clear how carefully these people construct the new sketchy password. Therefore, we asked these 28 participants in detail about their strategy of creating the new sketchy password. Their responses provided answers to our Research Question 1.

Table 3.2: Type of modifications made when reusing an existing password to create a sketchy password.

Modification	Never	Seldom	Sometimes	Often	Total
Add/delete 1-2 characters	4	5	8	6	23
Add/delete more than 2 characters	6	7	7	3	23
Replace 1-2 characters with others	4	5	7	7	23
Add special symbols	5	2	10	6	23

Reuse Without Modification Almost half of these participants (13 out of 28) said that when creating a new sketchy password, they “never” reuse a password they have used elsewhere without any modification. We compared participants who reported “never” to “seldom” reusing their identity or financial passwords against those that “sometimes” to “often” reused them. Overall, 24 of these 28 participants reported “never” to “seldom”, whereas only 4 participants reported “sometimes” to “often”, $\chi^2(1) = 6.63$, $p < .05$.

Reuse With Modification The reuse with modification rate is comparatively high among the participants. Only 5 out of 28 participants (18%) said that when constructing a new sketchy password, they “never” reuse a password they have used elsewhere, with some modifications. Specifically, 13 of these 28 participants reported “never” to “seldom” reusing their identity passwords (with some modifications), whereas 15 participants reported “sometimes” to “often” reusing them, $\chi^2(1) = 0$, $p = 1$. For financial passwords, 24 reported “never” to “seldom” reusing them (with some modifications), whereas 4 reported “sometimes” to “often” reusing them, $\chi^2(1) = 6.63$, $p < .05$.

These participants also answered what kind of modifications they make. Their responses are shown in Table 3.2.

Table 3.3: Type of similarity between an existing password and the new sketchy password.

Type of similarity	Never	Seldom	Sometimes	Often
Both are inspired by common source	4	7	10	6
Both are dictionary words/minor variations	11	10	5	1
Both are English phrases	10	7	5	5
Both are related to friend/family	11	5	8	3
Both are personally meaningful words	6	5	10	6
Both are personally meaningful numbers	6	5	9	7

Reuse Similar Process These 28 participants also indicated that they use a similar process as they have used before when creating a new sketchy password. Only a single participant (4%) reported that he/she “never” uses a similar process. In particular, 18 of these 28 participants reported “never” to “seldom” using a similar thought process of constructing an identity password when constructing a sketchy password, while 10 participants reported “sometimes” to “often” doing so, $\chi^2(1) = .66, p = .42$. On the other hand, 27 of these 28 participants reported “never” to “seldom” using a similar thought process of constructing a financial password when constructing a sketchy password, while only a single participant reported “sometimes” to “often” doing so, $\chi^2(1) = 13.11, p < .001$.

Finally, we asked these participants how the new sketchy password would be similar to their existing passwords. Their responses are shown in Table 3.3.

We consider two types of similarity: semantic similarity and syntactic similarity. If two passwords are similar based on their meaning or semantic content, we denote this as semantic similarity. For example, two passwords are semantically similar if they are inspired from a common source (literature, film, music etc.) or if both are personally meaningful words or numbers. On the other hand, two passwords are syntactically similar if one is a minor variation of another. Examples of minor

variation are capitalizing a letter, adding digits/special characters, replacing one letter with another etc.

Option 2 referred to syntactic similarity (both passwords are dictionary words, or minor variations of those words), whereas options 1, 5 and 6 were related to semantic similarity. The responses of the participants clearly showed that they follow options 1, 5 and 6 more frequently than option 2 (Table 3.3). This demonstrated that the semantic similarity is much more evident than the syntactic similarity. This issue is discussed further in Section 3.5.2.

3.4.4 Content Password

We report our results about content passwords by relating them to Hypothesis 2.b.

Hypothesis 2.b

As suggested in the first part of Hypothesis 2.b, it was found that most of the users use a fixed set of passwords for sketchy sites. Specifically, 66 of our 80 participants (83%) reported using a fixed set of passwords for content sites whereas 14 reported that they create a new password each time they open a new account at a familiar content site. The chi-square statistics demonstrated that the difference was significant, $\chi^2(1) = 17.47, p < .001$.

For the second part of the Hypothesis 2.b, it was proposed that of the participants that do use this fixed set of passwords that they usually do not reuse this fixed set of passwords in their important financial or identity accounts. To assess this, we compared participants who reported “never” to “seldom” reusing against those that “sometimes” to “often” reused. This part of the hypothesis was not supported, 31 of these 66 participants reported “never” to “seldom”, whereas 35 participants reported

“sometimes” to “often” reusing these fixed passwords for their financial or identity accounts, $\chi^2(1) = .03$, $p = 0.863$.

Thus, Hypothesis 2.b turned out to be partially true.

Only a minority of participants (18%) said that they construct a new password each time they create a new account in any content site. We do not report in detail on their strategies for creating the new content passwords, but their responses were consistent with the responses for sketchy passwords. For example, the responses also suggested that the semantic similarity is much more evident than the syntactic similarity.

3.4.5 Password Sharing

The responses of the participants showed that almost all of the participants share passwords with others for maintaining a shared account. Only 6 out of 80 participants (8%) reported that they “never” share a password with others for maintaining a shared account.

We compared participants who reported “never” to “seldom” sharing against those that “sometimes” to “often” shared. Only 24 of our 80 participants reported “never” to “seldom”, whereas 56 participants reported “sometimes” to “often”, $\chi^2(1) = 5.86$, $p < .05$.

Research Question 2

We asked the 74 participants (participants who reported “seldom”, “sometimes”, or “often” sharing) in detail about their strategy of creating a shared password. Their responses provided answers to Research Question 2.

Reuse Without Modification We first asked them the extent to which they create a shared password by reusing an identity password they have used elsewhere, without any modification. We compared participants who reported “never” to “seldom” reusing against those that “sometimes” to “often” reused them. Overall, 43 of these 74 participants reported “never” to “seldom”, while 31 participants reported “sometimes” to “often”, $\chi^2(1) = .68, p = .409$.

We also asked them the extent to which they create a shared password by reusing a financial password they have used elsewhere, without any modification. Overall, 58 of these 74 participants reported “never” to “seldom”, while only 16 participants reported “sometimes” to “often”, $\chi^2(1) = 11.76, p < .001$.

Reuse With Modification We next asked these participants the extent to which they create a shared password by reusing an identity password they have used elsewhere, with some modifications. Among 74 participants, 40 reported “never” to “seldom”, whereas 34 reported “sometimes” to “often”, $\chi^2(1) = .11, p = .74$. For financial passwords, 59 out of these 74 participants reported “never” to “seldom”, whereas only 15 reported “sometimes” to “often”, $\chi^2(1) = 13.07, p < .001$.

These participants also answered what kind of modifications they make. Their responses are shown in Table 3.4.

Table 3.4: Type of modifications made when reusing an existing password to create a shared password.

Modification	Never	Seldom	Sometimes	Often	Total
Add/delete 1-2 characters	8	13	22	19	62
Add/delete more than 2 characters	15	18	19	10	62
Replace 1-2 characters with others	6	10	28	18	62
Add special symbols	20	8	20	14	62

Table 3.5: Type of similarity between an existing password and the new shared password.

Type of similarity	Never	Seldom	Sometimes	Often
Both are inspired by common source	11	9	23	24
Both are dictionary words/minor variations	30	25	8	4
Both are English phrases	22	16	14	14
Both are related to friend/family	22	10	17	17
Both are personally meaningful words	9	5	27	26
Both are personally meaningful numbers	10	8	23	26

Reuse Similar Process The responses of these 74 participants indicated that they use a thought process similar to the one used for creating an identity password when they create a shared password. Specifically, only 29 of these 74 participants reported “never” to “seldom”, whereas 45 participants reported “sometimes” to “often”, $\chi^2(1) = 1.34$, $p = .247$. For financial passwords, however, 57 out of these 74 participants reported “never” to “seldom”, whereas only 17 reported “sometimes” to “often”, $\chi^2(1) = 10.53$, $p < .05$.

Finally, we asked these participants how the new shared password would be similar to their existing passwords. Their responses are summarized in Table 3.5. We can see that the semantic similarity is much more evident than the syntactic similarity in this case as well (participants follow options 1, 5 and 6 more frequently than option 2).

3.4.6 Password Cracking

We did not only rely on what the participants said during the survey in Phase 2, we also analyzed on the basis of what they did in Phase 1. We exploited the passwords constructed by the participants in Phase 1 and tried to crack the financial and identity (higher-level) passwords of a participant by using that participant’s content

and sketchy (lower-level) passwords. For cracking purposes, we used the John The Ripper (JTR) password cracker.

Attack Model We first assume a scenario where an attacker compromises one lower-level password of each participant. We calculate the percentage of higher-level passwords that could be cracked by the attacker under this assumption. We also observe the effect when each additional lower-level password is compromised by the attacker. Specifically, we try to answer the following questions:

Question 1 What percentage of higher-level passwords could be cracked by an attacker by compromising one lower-level password of each participant?

Question 2 With compromise of each additional lower-level password, what additional percentage of higher-level passwords could be cracked by the attacker?

Attack without wordlist In this attack mode, we tried to crack the higher-level passwords of a participant by using the lower-level passwords only, without using any wordlist or dictionary. We performed cracking by using JTR in our modified wordlist mode (Section 3.3.4). We used the word mangling rules of JTR to mangle the lower-level passwords in order to guess the higher-level ones. These mangling rules include appending digits, replacing letters with similar symbols (\$ instead of S, for example) etc. For each participant, the wordlist consisted only the lower-level passwords of the same participant.

Figure 3.7 demonstrates the password cracking statistics. By using one lower-level password of each participant, we could successfully crack 10.6% of higher-level passwords. The percentage increased to 19.1% (61 out of 320) when we used all the four lower-level passwords to crack the higher-level ones.

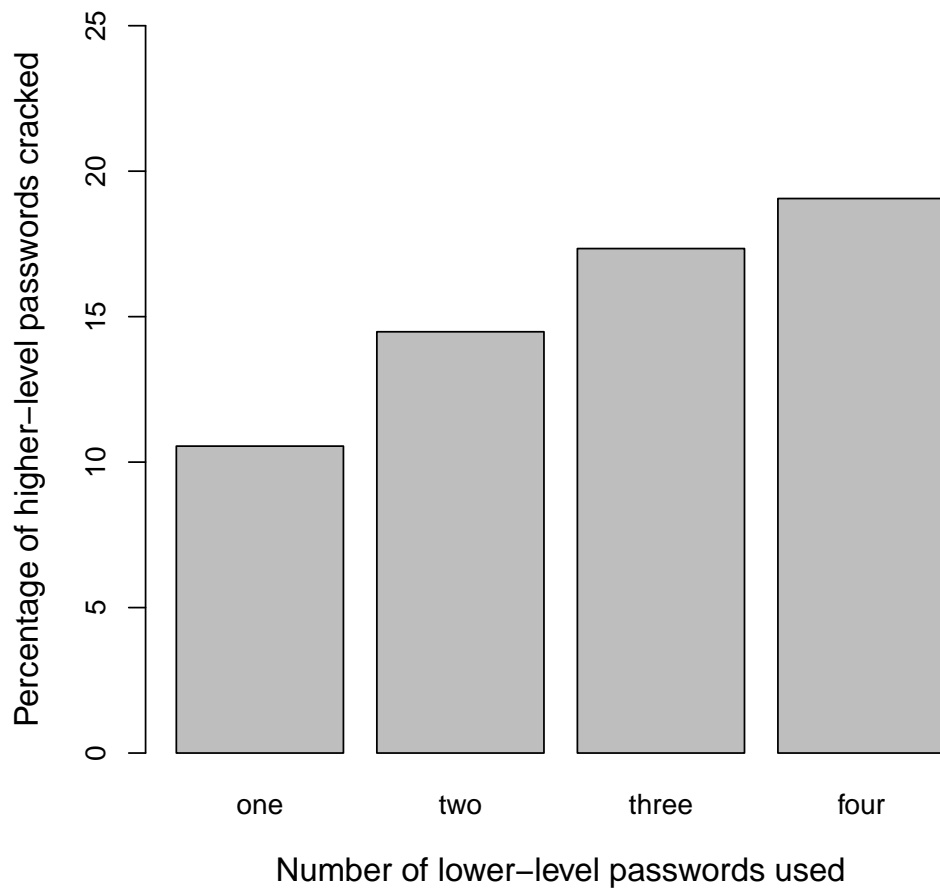


Figure 3.7: Password cracking statistics (without wordlist).

Less than half of the cracked higher-level passwords (29 out of 61) were the same as the lower-level ones. The rest were minor modifications of the lower-level passwords, such as appending digits, appending years, capitalizing first/middle letter, etc. Thus, even when a password is not directly reused, reuse with modification greatly increases the risks to users. Figure 3.8 demonstrates the password reuse mechanism of the participants.

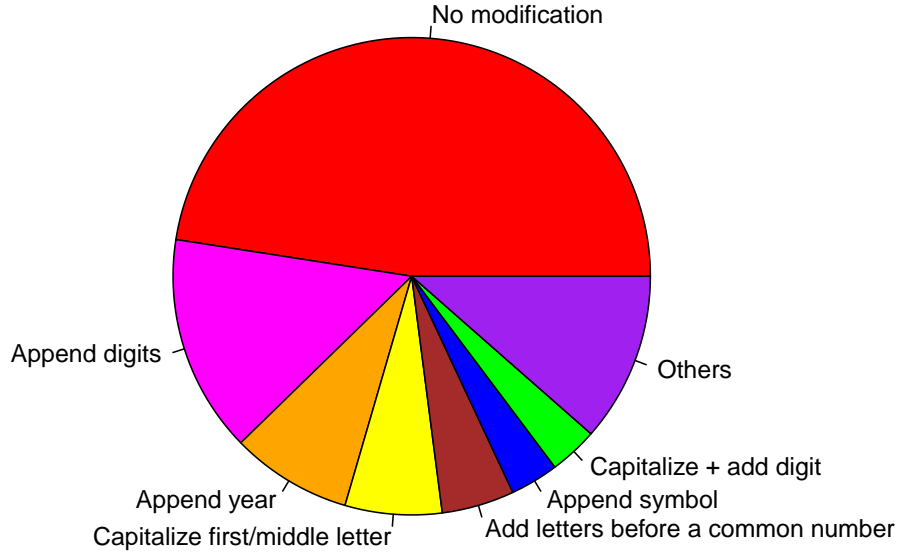


Figure 3.8: Password reuse mechanism.

Attack with wordlist In this attack mode, we impersonated a more sophisticated attacker who would use a wordlist along with the lower-level passwords for performing the cracking operations. For each participant, we combined the participant’s lower-level passwords with the Cain & Abel wordlist and tried to crack the higher-level passwords of the same participant by using JTR in our modified wordlist mode.

Figure 3.9 demonstrates the password cracking statistics. By using Cain & Abel wordlist and one lower-level password of each participant, we could successfully crack 26.8% of higher-level passwords. The percentage increased to 33.1% (106 out of 320)

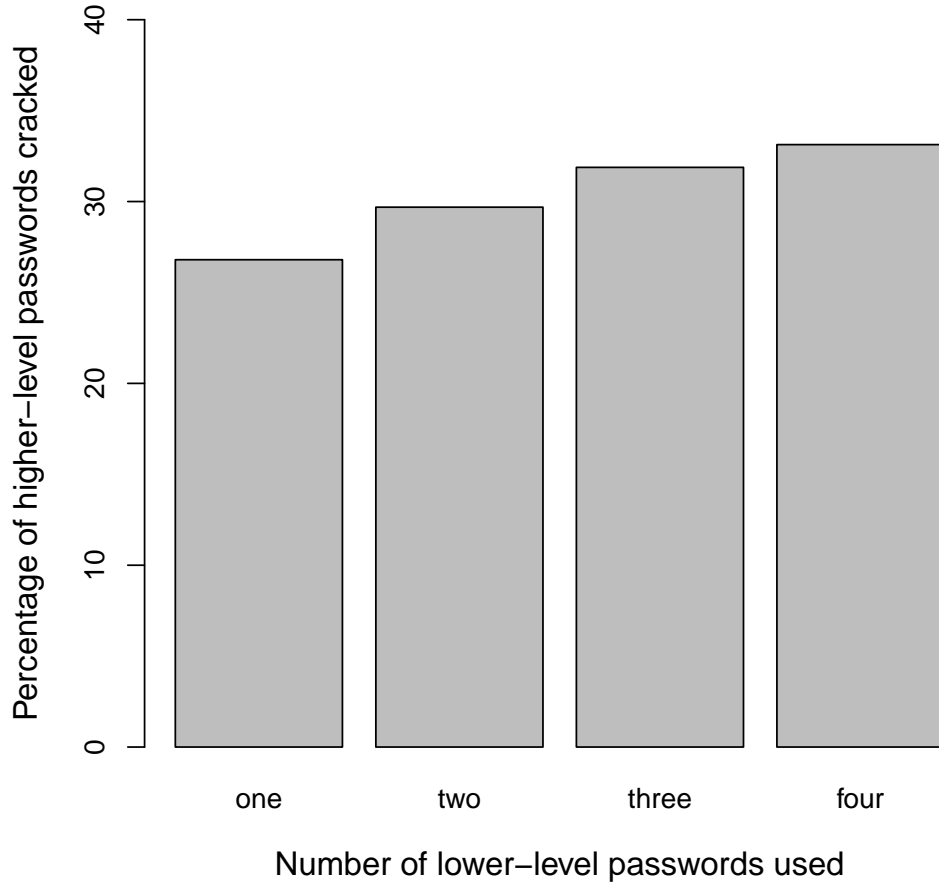


Figure 3.9: Password cracking statistics (with wordlist).

when we used all the four lower-level passwords along with the Cain & Abel wordlist to crack the higher-level passwords.

We also tried to crack the higher-level passwords of the participants by using the Cain & Abel wordlist (in our modified wordlist mode) only, without using the lower-level passwords. Among the 320 higher-level passwords, we could successfully crack 21.9% (70 out of 320) of passwords with this method. Thus, by combining the lower-level passwords with the Cain & Abel wordlist, we could successfully crack

51.4% more higher-level passwords (the number of cracked higher-level passwords increased from 70 to 106). A chi-square test showed that the difference is significant, $\chi^2(1) = 9.6, p < .05$.

3.5 Discussion

Before discussing the various implications of our findings, it is important to highlight several limitations of our study.

3.5.1 Limitations

First, we do not dispute the fact that it is difficult to demonstrate ecological validity [14] in any password study where participants are aware that they are creating passwords for experimental purpose, rather than for accounts they value in real life for regular use over a long period of time. However, in the context of this work, passwords for websites of different categories were created under conditions that should be affected equally by this issue. Furthermore, prior work suggested that involving a role-play scenario would motivate users to construct passwords more seriously than a survey scenario [68]. We thus presented a role-play scenario for each website, and the sites were designed so that they would resemble real-world sites as much as possible (Section 3.3.2.1). We note, however, that our participants were not required to return on a second day to re-enter their passwords, and as such, some of them might have constructed less memorable random passwords.

It is also difficult to emulate the temporal, situational, environmental and other real-life contextual aspects of password construction in a laboratory study. For example, when constructing a password at home, a user with a pet on his/her lap might construct a password that has semantic relation to that pet. However, the results of

a recent work of Fahl et al. on the ecological validity of password study reveal that passwords collected during user studies closely resemble users' actual passwords [35].

We also agree that the sample size of our study is not very large ($n=80$) and the study is only of university students, who may vary significantly from other populations in their password behavior, and in particular their password sharing behavior. However, compared to the sample sizes of prior works [121, 83, 41], which were also laboratory experiments among students, our sample size can be considered reasonable. Finally, we note that the presence of an observer may, if anything, encourage users to create stronger passwords than they might otherwise. This notion was supported by the results of our pilot study.

In this work, the categories of websites have been proposed by us based on prior work and our certain observations, it did not originate from the users in the first place. During the first phase of the study, the participants constructed passwords for eight websites which already conformed to our categorization. During the second phase, the categories of sites were described to participants and they subsequently answered questions about them. It would be unfair to claim based on our current study results that users do indeed classify accounts (and passwords) in this way independently in their real-life. Some non-primed feedback techniques such as card sorting might further help us to better understand the actual thought processes of the users. We plan to work on this in future.

Our password cracking results should be interpreted with caution, particularly by considering the fact that the participants were asked to construct eight passwords in a row in a short laboratory study with an artificial setting. It was a cognitively demanding task which might have prompted some of the participants to construct a bunch of similar (if not exactly the same) passwords.

We also acknowledge the fact that different password policies of the selected websites precluded us from making a fair comparison among the categories. The trade-off here was between realism and learning about user behavior. We chose the more realistic approach (enforced exactly the same password policies as they are enforced in the real sites). However, Section 3.4.2.1 contains the results of some fair comparisons and they are consistent with the basic claim of Hypothesis 1.

3.5.2 Implications

One important finding of our study is that users mentally classify their passwords into different categories according to the perceived importance level of the site. Our results suggested that users consider their financial passwords to be the most important. In our analysis, financial passwords were perceived to be significantly more important than identity passwords. Identity passwords, in turn, were perceived to be significantly more important than content and sketchy passwords. The perceived importance of passwords did not differ significantly between webmail accounts and social networking accounts, the two types of identity account that we studied.

Although the participants considered content passwords to be less important, their responses indicated that these passwords have a strong degree of similarity with their important financial and identity passwords. More than four-fifths of the participants reported that they use a fixed set of passwords for all kinds of content sites, which is a reasonable practice. More than half of these participants, however, further indicated that they reuse this fixed set of passwords in their important financial or identity accounts. Of those participants, about one-third reported that they “often” reuse them in identity sites, while one-fifth reported that they “often” reuse them in financial sites.

These findings suggest that a password used at a content site may be more valuable than the account which it protects. An account at a content site typically does not contain much sensitive information; users create it mainly for the purpose of customizing the site experience. The passwords used to protect these accounts, however, are valuable because they are reused frequently in identity or financial websites. If a content password is leaked, it can be used effectively to compromise important identity or financial accounts. This issue should be considered when formulating the authentication policies for content sites. Given how entrenched passwords are as an authentication mechanism, however, it may be more useful to help and encourage users at financial and identity sites to make stronger but memorable passwords that are clearly distinct from their content site passwords.

The passwords we collected from the first phase of the study showed that users construct stronger passwords (passwords of longer length with more capital letters, digits, and special characters) for higher-level sites. The rating of the eight websites also confirmed that users distinguish between higher-level sites and lower-level sites. However, the password cracking statistics and the responses of the survey in the second phase suggested that higher-level passwords have a good degree of similarity with lower-level passwords. Thus, it is apparent that while users do have a notion that sites have different levels of security and importance, expedience and simplicity of password management trump what they know are more secure behaviors.

Another important finding of our study is that the degree of semantic similarity is stronger than the syntactic similarity among passwords of different levels of a user. Our cracking methodology through JTR relied only on syntactic similarity. Through word mangling rules, it modified the lower-level passwords in various ways in order to guess the higher-level passwords. Semantic similarity was not examined or used. For example, multiple passwords of a user can be inspired from common source (e.g.,

music, film, sports etc.). If one of the passwords of a user is related to a personally meaningful word (e.g., the name of her cat), then it is likely that another of her passwords is also inspired by a similar thing (e.g., the name of her family’s cat from when she was a child). In fact, the users’ responses suggested that these practices are followed frequently (Table 3.5 and Table 3.3). Our cracking methodology did not leverage these kinds of semantic similarity. We believe that by exploiting semantic similarity, a larger percentage of higher-level passwords can be cracked. We leave this as a future work.

The issue of semantic similarity has a wider implication for shared passwords. Our survey responses showed that when creating a shared password, users frequently use a similar thought process as used when creating their identity (email/Facebook) passwords. This reveals a serious breach in the privacy of their identity accounts. As discussed before, users generally share accounts among close friends, family, or apartment mates. If the shared password has any kind of semantic similarity with her other passwords, then it becomes easier for these people to guess those other passwords.

Passwords for shared accounts are also frequently created by reusing important identity or financial passwords. More than half of the participants reported that they “sometimes” or “often” reuse a password without any modification when creating a password for a shared account. All of these findings highlight the indirect consequences of password sharing and suggest that password sharing perhaps should not be considered as a “nuanced practice engaged in with thought and care”, as suggested by Kaye in [64].

3.5.3 Conclusion

In this work, we proposed a hierarchy of users' Web passwords based on the perceived importance level of the sites and conducted a user survey to verify the hierarchy. The responses demonstrated that users consider their financial passwords to be significantly more important than their identity passwords, which, in turn, are considered to be significantly more important than their content and sketchy passwords.

We also conducted a laboratory experiment where we asked the participants to construct these four types of passwords. We exploited the content and sketchy (lower-level) passwords of a participant along with a password-cracking dictionary to crack that participant's identity and financial (higher-level) passwords. We could successfully crack almost one-third (106 out of 320) of the higher-level passwords in this method. This number is significantly higher than the number of passwords cracked by using the password-cracking dictionary only, without using the lower-level passwords.

This work also highlighted the indirect consequences of password sharing. In our survey, we asked our participants regarding their password construction practices for shared accounts (accounts for paid subscription sites like Netflix or accounts for sharing a common service such as wireless Wi-fi). We found out that users use a thought process similar to the one used for creating an identity password when they create a shared password. Thus, passwords for shared accounts could be exploited to compromise important identity accounts.

An attacker could also exploit the content passwords to compromise users' identity accounts. Our survey results revealed that although most of the users are conscious and use a fixed set of passwords for content sites, a majority of these conscious users further reuse this fixed set of passwords for their important identity accounts.

These findings show that although users consider their identity passwords to be significantly more important than their lower security level passwords, they are not conscious enough to protect themselves from attacks that might leverage these lower security level passwords to guess their identity passwords. For financial passwords, users are relatively more conscious. However, the percentage of users who reuse their financial passwords to construct their lower-level passwords is not nominal.

We acknowledge the fact that our hierarchy of Web passwords did not originate from the users. We proposed it based on prior work and certain observations, and later verified it by conducting a comprehensive user study. There is more to do to completely understand how users mentally classify all of their password-protected accounts in real life. When cracking the higher-level passwords, we also did not consider the semantic similarity at all. We plan to work on these issues in future.

CHAPTER 4

Towards Creating Stronger Passwords by Using Mobile Phone Handsets

In this chapter, we describe the two experiments which were conducted for examining the association between password strengths and the interface through which they are created. We also tested the effectiveness of a custom keypad layout designed by us for assisting users in creating stronger passwords when using mobile phone handsets.

4.1 Experiment 1

In February 2013, we conducted the first laboratory experiment with 72 UTA students (45 female and 27 male).

4.1.1 Study Administration

We administered the study through the *research pool* of the department of psychology at UTA. We also recruited a few participants (N=9) from outside the research pool. Those participants received a 5 dollar gift voucher for a local restaurant.

4.1.2 Apparatus

For our experiment, we used a Motorola MILESTONE A853 mobile handset running Android 2.1. This handset contains a slide-out physical keyboard and also a QWERTY-type touchscreen keypad.

We designed our custom touchscreen layout by adding two extra rows of characters on the screen, as shown in Figure 4.1. One row contained the ten digits and

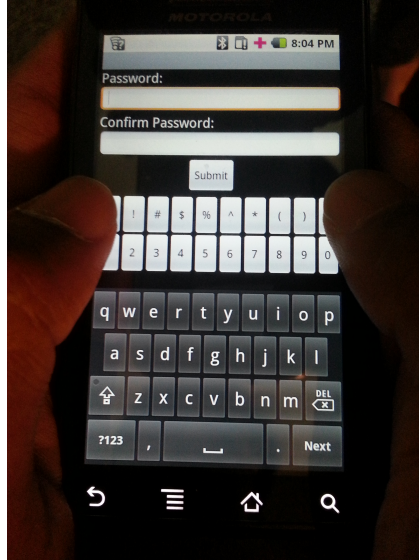


Figure 4.1: Custom layout with two extra rows of keys

the other row contained ten common special characters. These common special characters are the ten characters that appear along with the ten numeric keys on the second row of a standard desktop keyboard. The size and the inter-key distance of the additional keys were same as the original keys.

4.1.3 Experimental Groups

In our experiment, we asked the participants to construct new passwords. It was a between-group experiment and each participant constructed passwords by using one interface only. We randomly assigned each participant to one of the four groups:

- Computer keyboard (*keyboard* group)
- Mobile phone with physical keyboard (*physical* group)
- Mobile phone with touchscreen keypad (*touchscreen* group)
- Mobile phone with custom layout (*custom* group)

The *keyboard* group was provided with a standard desktop computer keyboard. The other three groups were given the Motorola MILESTONE handset, but they were presented with different layouts of the handset.

The *physical* group was asked to use the physical keyboard of the Motorola MILESTONE handset, while the *touchscreen* group was asked to use the touchscreen keypad of the same device. Since both the *physical* and *touchscreen* groups used the same device to construct passwords, confounding effects like the convenience of holding the device were removed. The *custom* group was asked to use our custom layout.

4.1.4 Password Construction

The participants from each group were asked to construct new passwords in their respective interfaces for two different banking websites: Chase.com and Wellsfargo.com. We wanted the participants to spontaneously construct secure passwords that would be relatively long and would contain digits, capital letters, and special characters. Therefore, we selected banking websites to trigger the sense that security is important among the participants without explicitly asking them to construct strong and secure passwords. We selected Chase.com and Wellsfargo.com because these two banks should be familiar to the participant students due to the prevalence of their ATMs on the university campus.

For ethical reasons and security purposes, we explicitly told the participants not to provide any of their existing passwords. For both websites, we provided a brief introduction and presented a real-life scenario to the participants. For Chase.com, the participants were presented with the following scenario:

Chase is one of the largest banks in the US and it has an ATM on campus. Imagine that you are creating an account at Chase.com for online banking. You have reached the final step of creating your new account, and you need to create a password. Proceed to the next page to input your new password.

When they clicked OK, the password construction page appeared. Once they constructed the password for Chase.com, a similar scenario was presented for Wells-fargo.com.

After constructing the two passwords, the participants were asked to answer a few questions about their mobile handsets. Demographic questions were asked at the end of the study.

4.1.5 Results

We calculated the mean entropy of the passwords for each of the four interfaces. The entropy was calculated by using the formula, entropy $H = L \log_2 N$, an approximation of plain Shannon entropy, where L is the length of the password and N is the size of the alphabet. The alphabet size is the sum of the sizes of different types of characters. These types and sizes are:

- Lowercase letters: 26
- Uppercase letters: 26
- Digits: 10
- Common special characters: 10
- Uncommon special characters: 22

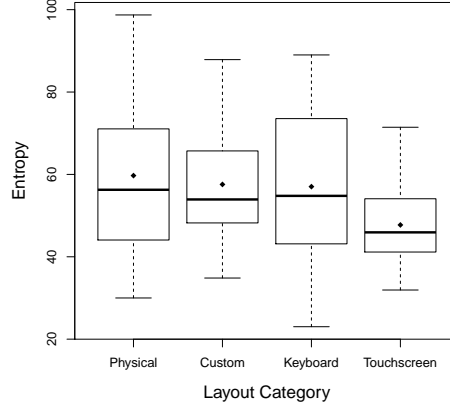


Figure 4.2: Box plot of entropy values. The mean values have been highlighted by the black points.

As mentioned before, the common special characters are the ten characters which we included in a separate row in our custom layout. We distinguished between common and uncommon special characters so that an addition of a common special character from the custom layout would not unduly increase the entropy of the password (increase the alphabet size by 10 instead of 32).

Figure 4.2 summarizes the entropy values for the four interfaces. As can be seen in Figure 4.2, The mean entropy is higher for *physical* group and then shorter for *custom*, *keyboard*, and *touchscreen* group in order. This finding goes beyond the general intuition that passwords that are constructed by using computer keyboards would be the strongest.

We conducted a one-way Anova test to analyze the differences between mean entropies for the four interfaces. A one-way Anova test is the standard way to analyze the differences between more than two mean values. It is basically a generalization of the t-test, where the number of groups are more than two, and helps to reduce the chance of incorrect findings of significance compared with multiple pairwise t-tests.

Table 4.1: Summary of Tukey’s post-hoc analysis. For each pair of interfaces, the difference, the 95% confidence interval and the p-value of the pairwise comparison are shown.

Pair	Diff	Lower	Upper	p-value
Custom-Keyboard	0.56	-9.37	10.48	0.999
Physical-Keyboard	2.90	-7.02	12.83	0.872
Touchscreen-Keyboard	-9.73	-19.65	0.20	0.057
Physical-Custom	2.35	-7.58	12.28	0.927
Touchscreen-Custom	-10.28	-20.21	-0.35	0.039
Touchscreen-Physical	-12.63	-22.56	-2.70	0.006

The results demonstrated that entropy of passwords differed significantly across the four interfaces, $F(3, 140) = 4.28$, $p < .01$.

Since the difference value was significant for Anova, we also conducted Tukey’s post-hoc comparisons to confirm where the differences occurred between groups. The results indicated that entropies were significantly higher for the *custom* group than those for the *touchscreen* group, $p < .05$. Also, the difference between the *physical* group and the *touchscreen* group was highly significant, $p < .01$. Table 4.1 summarizes the results for the Tukey’s post-hoc test.

We conducted another one-way Anova test to compare between keyboard (*physical* and *keyboard* groups combined) and keypad (*touchscreen* and *custom* groups combined) interfaces. As predicted, entropies were significantly higher for the keyboard interface than those for the keypad interface, $F(1, 142) = 4.80$, $p < .05$.

4.1.6 Discussion

The design and the results of the first experiment left room for an alternative explanation for better performance of *custom* and *physical* groups. The experimental results demonstrated that the *physical* group created stronger passwords than the *keyboard* group. One possible reason for this might be the fact that among the four

interfaces, computer keyboards are the most widespread. The participants of the *keyboard* group were already familiar with the interface and did not contemplate much while constructing their passwords. Since they did not face much difficulty in typing a password, they were relatively less engaged in their password construction activity. As a result, the mean entropy was lower for *keyboard* group than *physical* group.

This same bias would apply to the *custom* vs *touchscreen* group comparison since people are more familiar with the standard touchscreen than the custom touchscreen. Thus, the alternative explanation for the better performance of our custom layout could be phrased as: “If mobile layout designers really were to adopt the custom layout, would users would become accustomed to it, causing it to lose its advantage?”.

The reason for *physical* group’s better performance than the *touchscreen* group could also be explained in this way. Our post-experimental brief survey regarding handset usage confirmed that a majority of the participants (47 out of 72) primarily use mobile handsets that contain touchscreen keypads. Thus, our sample population for Experiment 1 was predominantly familiar with touchscreen keypads, resulting in the same potential bias for *physical* vs. *touchscreen* group.

Experiment 2 was designed to address both of these alternative explanations by adding artificial tasks that required the participants to become accustomed to the interfaces before getting to the password creation task. Before creating passwords, participants were asked to complete other formalities of creating a new bank account. This, in turn, ensured that all the participants were already accustomed to the interfaces before creating the passwords, and allowed for a more fair comparison. A supplementary feature of Experiment 2 is that it was designed as a within-group experiment where each participant was asked to construct passwords by using two

interfaces, which allowed for a more straightforward comparison between the interfaces.

4.2 Experiment 2

In June 2013, we conducted a second laboratory experiment with 24 students (14 female and 10 male). As with Experiment 1, we recruited participants from the research pool of UTA and used the Motorola MILESTONE A853 mobile handset.

4.2.1 Password Construction

In Experiment 2, we exclusively focused on comparing the standard touchscreen with the two other mobile phone layouts: physical keyboard layout and our custom touchscreen layout. This yielded two experimental groups and each participant was randomly assigned to one of the groups:

- Standard touchscreen vs. custom touchscreen
- Standard touchscreen vs. physical keyboard

4.2.1.1 Standard Touchscreen vs. Custom Touchscreen

The participants in this group were first presented with the following instructions:

Chase is one of the largest banks in the US and it has an ATM on campus.

Imagine that you are creating an account at Chase.com for online banking.

Proceed to the next page to start creating your new bank account.

When a participant clicked OK, she was presented with a set of artificial tasks to be completed using the first layout she had been assigned. The tasks were designed in such ways that they would resemble the usual steps of creating a new bank

account. The tasks involved entering assigned dummy values, written on a piece of paper, for name, account number, address, and email address. The assigned address contained multiple special characters. Thus, while typing these dummy values, the participants got accustomed to typing capital letters, digits, and special characters in their respective interfaces.

After entering these dummy values, participants were asked to answer some questions like “How much daily withdrawal limit do you want?”. Finally, they were redirected to the password construction page¹. Once the password was constructed, the following message was displayed:

Thank you for registering a new online account with Chase.com. For getting the full benefit of our online banking, we would like you to enroll in our ChaseQuickPay service. With Chase person-to-person QuickPay service, you can send money freely to anyone using their email address or mobile number.

Proceed to the next page to start the enrolling procedure in ChaseQuick-Pay service.

When the participants clicked OK, additional artificial tasks were provided, this time using the second assigned layout. As with the Chase account, the participants were asked to enter the same dummy name, address, and email address, plus a different account number (called the ChaseQuickPay ID). They were also asked to answer a few questions like “How much daily transfer limit do you want?”. After completing all these steps, they were redirected to the password construction page for the Chase-QuickPay service and specifically asked to construct a new password that would be different from the previous Chase bank account password.

¹As before, the participants were asked not to provide any of their existing passwords.

We randomized the order of presenting the layout to the participants. Thus, half of the participants constructed the Chase password by using the standard layout and the ChaseQuickPay password by using the custom layout. The remaining half followed the opposite order.

4.2.1.2 Standard Touchscreen vs. Physical Keyboard

Note that the participants of the previous group did not require switching the interface since both of the layouts were identical except the presence/absence of two additional rows of digits and special characters. However, the participants of this group were required to switch between the keypad and keyboard layout of the same handset.

To provide a plausible cover story for switching the interface in the middle of the experiment, we intentionally disabled the OK button when the ChaseQuickPay service message was displayed. As a result, participants were unable to proceed to the next step. At this point, the experimenter manually intervened and took the device from the participant. The experimenter pretended that the system had frozen for that interface, apologized to the participant, and asked her to complete the ChaseQuickPay registration formalities in the second interface. The post-experimental debriefing session showed that only a single participant could realize that this was an experimental manipulation for providing a plausible reason for switching the interface in the middle of the experiment. All the other procedures followed by this group were identical to those followed by the previous group.

4.2.2 Results

We used a paired t-test to compare the entropy values in standard touchscreen and custom touchscreen conditions. The results showed that entropy values were sig-

nificantly higher for custom touchscreen condition ($M=53.88$, $SD=7.75$) than standard touchscreen condition ($M=45.77$, $SD=9.69$); $t(11)=2.45$, $p=0.03$.

We also used a paired t-test to compare the entropy values in standard touchscreen and physical keyboard conditions. There was no significant difference in the scores for standard touchscreen ($M=42.21$, $SD=7.38$) and physical keyboard ($M=45.67$, $SD=11.98$) conditions; $t(11)=1.02$, $p=0.33$.

Finally, we carried out a paired t-test to compare the number of digits and special characters used by the participants in standard touchscreen and custom touchscreen conditions. The numbers were significantly higher for custom touchscreen condition ($M=3.92$, $SD=1.38$) than standard touchscreen condition ($M=2.33$, $SD=1.67$); $t(11)=2.78$, $p=0.02$. This confirms that our custom layout primed the participants to use more digits and special characters in their passwords.

4.2.3 Discussion

As with Experiment 1, our custom layout resulted in creation of passwords with significantly higher entropy values. The custom layout was introduced so that the participants could enter digits and special characters in a more convenient way. We predicted that passwords constructed by the custom layout would contain more digits and special characters than those constructed by the standard layout, which was in fact the result.

On the other hand, entropy values did not differ significantly between standard touchscreen and physical keyboard conditions. This indicates that the advantage of physical keyboard over standard touchscreen in Experiment 1 might have been an artifact of the design methodology. In Experiment 1, the participants who used the physical keyboard layout (*physical* group) were much more engaged during the password construction period since they were exploring a relatively less frequently used

layout. In contrast, by the time the participants had reached the step of constructing a password by using a physical keyboard in Experiment 2, they were already familiar with the layout. As a result, they were relatively less engaged during the password construction period, which might have resulted in creation of passwords with lower entropies.

To further validate this, we performed a cross-experiment entropy comparison. The results showed that while the mean entropy did not vary much between *touchscreen* group of Experiment 1 ($M=48.26$) and standard touchscreen conditions of Experiment 2 ($M=44$), it reduced drastically from *physical* group of Experiment 1 ($M=59.89$) to physical keyboard condition of Experiment 2 ($M=45.67$). An unpaired t-test showed a significant difference in the scores; $t(46) = 2.31$, $p = 0.025$. This suggests that user engagement is an important issue that affects the password construction process.

Two issues are worth mentioning regarding the methodology of Experiment 2. First, participants of the second group were required to switch interfaces in the middle of the experiment, and this may have impacted their performances. However, we note that half of the participants switched from touchscreen keypad to physical keyboard, while the remaining half switched from physical keyboard to touchscreen keypad. Thus, the effect of switching affected both of the interfaces equally. Secondly, we did not conduct a custom vs. physical group comparison in Experiment 2. As the main objective of Experiment 2 was to explore the potential bias effect of Experiment 1, and since the results of the touchscreen vs. physical group comparison confirmed the existence of this bias, we deemed custom vs. physical group comparison unnecessary. We also wanted to make a straightforward comparison between our custom layout and the standard layout.

4.3 General Discussion

In this section, we first discuss the ecological validity of our study. Next we discuss about the limitations of our study and implications of our findings. We also shed light on future research directions.

4.3.1 Ecological Validity

We do not dispute the fact that it is difficult to demonstrate ecological validity [14] in any password study where participants are aware that they are creating passwords for an experimental purpose, rather than for accounts that they value in real life for long-term use. Indeed, as we did not ask the participants in our study to return on a second day to re-enter their passwords, they were aware of the fact that there would be no consequences of their password choices. This might have impacted their password selection, since they had little incentive to create more memorable, and thus less secure, passwords.

However, we note that in our study, passwords for layouts of different categories were created under conditions that should be affected equally by this issue. Thus, we believe that the lack of a memorability test is not a critical issue for our study. Finally, we note that involving a role-play scenario as opposed to a survey scenario motivates users to construct passwords more seriously [68].

4.3.2 Low Sample Size

We acknowledge the fact that the sample size of our experiment was not very large. For this reason, any attempt to generalize our findings to the broader community should be made with care.

4.3.3 Standard Custom Layout

We relinquish any claim regarding the completeness of our proposed custom layout. Our custom layout should not be considered as a standard one. We designed the layout only to observe the behavior of the participants when they are presented with a more convenient option of inserting digits and special characters. Our results showed that the layout primed users to use more digits and special characters in their passwords, which in turn resulted in passwords with higher entropy values. Further research should be conducted before proposing any standard custom layout for mobile phones with touchscreen keypads.

4.3.4 Usability Evaluation

Our proposed custom layout has one limitation: it blocks the bottom part of the mobile phone screen. We note, however, that most existing password construction and password entry pages leave a considerable amount of blank space at the bottom of the page. Our custom layout, therefore, should not block any important portion of such pages during password entry. We plan to conduct a detailed usability study of our custom layout in the future.

4.3.5 Password Strength Measurement

We are aware of the fact that entropy is not the most appropriate measure of password strength [118]. For our studies, however, we mainly seek to capture how users' password behavior is constrained by keyboard layouts. The use (or lack thereof) of a variety of character types is captured reasonably well by our approximation of Shannon's entropy. Thus, we believe that entropy is more appropriate than measures such as difficulty of cracking, which is more dependent on the exact password choices of users than their ability to enter different types of passwords.

CHAPTER 5

Applying Psychometrics to Measure User Comfort when Constructing a Strong Password

5.1 Psychometrics

Psychometrics is the study of measuring complex psychological concepts, or *constructs*, such as a person's motivation, anger, personality, intelligence, attachment, or fear [84]. Since a construct is not a concrete material in the visible world, measuring a construct is not a straightforward task. For example, we know how anger looks, but we cannot describe in meters or grams how much anger a person feels. Psychometrics provides guidance to systematically develop and test a scale to measure this kind of psychological construct. In psychometrics, the basic component of a scale is referred to as an *item*. Items can be questions, true-false statements, or rating scales.

Although the field of psychometrics has been developed for measuring psychological constructs, we observe that the techniques of psychometrics may be suitable for other abstract constructs that concern human feelings and performance. The core function of psychometrics is to assign numbers to observations in a way that best allows people to summarize the observations. In other words, it tries to measure the psychological construct in a meaningful and interpretable way. Since usability is also an abstract construct [79], we believe that the techniques of psychometrics would be helpful in measuring the usability features of a security system in a meaningful and interpretable way.

5.1.1 Reliability and Validity

Let C be an arbitrary construct, such as happiness. At any given point in time, a person has a true level of happiness, namely X_T . A psychometric scale developed for measuring happiness, if administered on that person, will produce an observed level of happiness, namely X_O . The core job of a psychometrician is to develop a scale that produces a score X_O that approximates X_T as closely as possible.

The relationship between X_T and X_O can be formulated in this way [25]:

$$X_T = X_O + X_S + X_R, \quad (5.1)$$

where X_S comes from systematic sources of error and X_R comes from random sources of error. X_S refers to errors resulting from underlying stable characteristics of the construct, while X_R refers to errors that result from transient personal factors.

A scale may be characterized by two properties: reliability and validity. Reliability is the degree to which a scale produces stable and consistent results, and high reliability is indicated by low values of X_R (low random error). For example, if a person measures the weight of a penny several times by the same scale and always receives the same result, the scale is reliable. Validity is the degree to which a scale measures what it is purported to measure, and high validity is indicated by low values of both X_R and X_S (both low random and systematic errors). Note that a reliable scale may not be valid, such as if the scale consistently indicates that the penny weighs 100 kg., it suffers from high systematic error.

Although there are many validity classifications, one of the most prevalent frameworks recommends assessing validity from three perspectives: *content validity*, *construct validity*, and *criterion-related validity* [85, 47, 16]. Content validity refers to the extent that a scale represents a given construct, i.e. the extent to which the

content domain of the construct is represented in its entirety, and also the extent that items in the scale only represent the construct of interest. Construct validity refers to the extent that a scale assesses the underlying construct it is supposed to assess, i.e. whether the scale is accurately focused. Criterion-related validity, on the other hand, is the degree to which a scale score predicts meaningful outcomes in a real-life situation.

A sound psychometric scale should be reliable and valid in all three ways to have any meaningful application. As pointed out by Nunnally, however, validity is not an all-or-nothing property, rather it is a matter of degree [85].

5.1.2 Framework

The appropriate steps for developing a scale ultimately depend on the construct. Psychometricians have to know a number of tools and methodologies and have a thorough understanding of the construct to be measured so as to find the best mechanisms for developing and assessing the efficacy of the intended scale. There is substantial debate in the field in regards to the specific steps to employ to ensure the highest levels of validity. For example, marketing researchers often focus solely on differences due to stimuli changes, whereas psychology researchers are oftentimes interested in individual differences [25, 86]. However, there appears to be a consensus that comprehensive efforts that employ techniques from numerous perspectives are the most effective. We thus sought recommendations from various sources and applied heuristics from both marketing and psychology perspectives.

Our work is primarily based on the approach outlined by Nunnally in his various books [84, 85, 86]¹, but we also considered the recommendations provided by other no-

¹Nunnally’s seminal book “Psychometric theory”, published in 1967, had been widely used as the primary textbook in basic psychometric courses [84]. Eleven years later, he published a second

table psychometricians and statisticians, including Churchill [25], Parasuraman [89], and Kaiser [62].

5.2 Scale Development Steps

We now describe all the steps of our scale development procedure. We use the terms “layout” and “interface” interchangeably in the remainder of this chapter. For performing some of the statistical calculations, we used R packages such as `psych` and `nFactors`.

5.2.1 Domain Specification and Initial Item Pool Generation

The first step in developing a scale for measuring a construct is to specify the domain of the construct. A researcher must understand the construct thoroughly and determine its scope: what to be included and what to be excluded [25]. For example, in the context of our current work, “comfort of constructing a password” and “comfort of constructing a strong password” are two different constructs. The former mainly refers to general typing experience and may undervalue issues like “How easy is it to insert a special character in this layout?”, which is an important consideration for a strong password.

Churchill recommends performing a literature search and an experience survey for specifying the domain of the construct and generating the initial item pool [25]. An experience survey involves consulting a group of people who are considered to be knowledgeable in the domain. We conducted such a survey by forming a panel of two password researchers and two mobile UI specialists. We also consulted with expert researchers from marketing and psychology to obtain more substantive insights about edition by incorporating the new ideas that had been introduced over the decade [85]. He also co-authored a book with Bernstein, a notable clinical psychologist [86].

the scale development procedures. A one-on-one session was held with each of the panel members.

The marketing expert recommended to review the existing scales that have been developed to measure user engagement or customer satisfaction for various activities performed with a computer or mobile phone (online shopping, for example). The psychology expert suggested that we consider emotional or cognitive hindrances such as frustration or confusion that might affect the password construction activity. The mobile UI specialists recommended that we consider subtle typing issues, such as key sensitivity and inter-key distance, which are associated with the user experience when typing on a particular layout. The password researchers focused more on entropy and were interested to observe how different keypad layouts would affect the frequency of using capital letters, digits, and special characters when constructing a new password by using those layouts.

After consulting with all the panel members and reviewing the relevant literature, we generated an initial pool of 32 items.

The first set of items was developed to assess the ease of using a specific layout to construct a strong password. The strength of a password is associated with its length and the frequency of uppercase letters, digits, and special characters (see Section 5.5.2 for more discussion about password strength). Items in this category directly focused on assessing how easily a user could type an uppercase letter and insert digits and special characters by using a specific layout. We also conjecture that a user would be motivated to type a longer password if her general typing experience is good when using a specific layout. Thus, we tried to capture the general typing experience of a user through this set of items. Accordingly, items in this category focused on issues like ease of editing, key sensitivity, and inter-key distance.

1. It was easy to type an uppercase letter in this layout.
2. It was easy to insert a numeric digit in this layout.
3. It was easy to insert a special character in this layout.
4. It was easy, overall, to type passwords using this layout.
5. I could easily type the exact letter that I wanted to type in this layout.
6. The distance between the keys was not very close in this layout.
7. The keypad of this layout was too much sensitive to my touch.
8. The keypad of this layout was too little sensitive to my touch.
9. It was easy to make edits when typing in this layout.
10. The keys were marked with familiar symbols in this layout.
11. I could clearly see the keys in this layout.
12. It was easy to type using both hands in this layout.

As pointed out by the psychology expert, emotional and cognitive hindrances might adversely affect the password construction activity of a user. The second set of items reflected this direction and were written as reverse-coded items [78]. Consequently, the wording of the items reflected negative connotations such as “annoyance”, “error”, “confusion”, and “restriction”.

13. I felt annoyed when typing an uppercase letter in this layout.
14. I felt annoyed when inserting a numeric digit in this layout.
15. I felt annoyed when inserting a special character in this layout.
16. I felt frustrated, overall, when typing passwords using this layout.
17. I made more errors in this layout when typing.
18. It was confusing trying to find some keys in this layout.
19. I found this layout confusing to use when I was typing an uppercase letter.
20. I found this layout confusing to use when I was inserting a numeric digit.

21. I found this layout confusing to use when I was inserting a special character.
22. The current method of typing an uppercase letter in this layout restricted me from using more uppercase letters in my passwords.
23. The current method of inserting a numeric digit in this layout restricted me from using more digits in my passwords.
24. The current method of inserting a special character in this layout restricted me from using more special characters in my passwords.
25. The current method of typing in this layout restricted me from typing a longer password.

The final set of items targeted user satisfaction. Items in this category addressed whether the users felt that there should be an easier way to insert digits or special characters, whether they were able to type quickly by using the layout, and so on.

26. I want an easier method of typing an uppercase letter in this layout.
27. I want an easier method of inserting a numeric digit in this layout.
28. I want an easier method of inserting a special character in this layout.
29. I was able to quickly type an uppercase letter in this layout.
30. I was able to quickly insert a numeric digit in this layout.
31. I was able to quickly insert a special character in this layout.
32. I was able to quickly type passwords using this layout.

5.2.2 Content Validity Assessment

After generating the initial item pool, the items were subjected to an assessment of content validity. As mentioned before, content validity refers to the extent to which a scale represents the content domain of a construct [46]. Content validity should be assessed immediately after developing the items, as this provides an opportunity to

refine the items before making large investments in administering the items to a sample population [104, 98].

We assessed the content validity of each item by following Lawshe’s guidelines [71]. Lawshe proposes forming a panel of subject matter experts and asking each of them to rate each item in terms of whether the knowledge or skills measured by that item is “essential”, “useful, but not essential”, or “not necessary” to the performance of measuring the construct. He developed a formula for measuring the content validity of each item [71]:

$$CVR = \frac{(n_e - \frac{N}{2})}{\frac{N}{2}}, \quad (5.2)$$

where *CVR* stands for *content validity ratio*, n_e is the number of panelists indicating that the item is “essential”, and N is the total number of panelists. Lawshe also provides a table of critical values of CVR for a given size of subject matter expert panel [71]. According to his recommendation, an item can be retained if its CVR value exceeds the critical value. Accordingly, we formed a panel of eight subject matter experts and asked them to evaluate our initial set of items. We recruited mobile application developers with at least two years of experience in working with mobile UI as our subject matter experts. They were explained beforehand about the purpose of the scale and the association between a strong password and a particular layout.

Out of 32 items, 19 items were retained (see Table 5.1 for the list of retained items), as their CVR was higher than the 0.75 threshold recommended in Lawshe’s table for a panel of eight subject matter experts. Two psychometricians reviewed the wordings of the retained items to avoid ambiguity.

Table 5.1: Reliability Analysis. Cronbach's α value is 0.96.

Item	Value
1. It was easy to type an uppercase letter in this layout.	0.76
2. It was easy to insert a numeric digit in this layout.	0.80
3. It was easy to insert a special character in this layout.	0.83
4. It was easy, overall, to type passwords using this layout.	0.90
5. I felt annoyed when typing an uppercase letter in this layout.	0.73
6. I felt annoyed when inserting a digit in this layout.	0.77
7. I felt annoyed when inserting a special character in this layout.	0.75
8. I felt frustrated, overall, when typing passwords using this layout.	0.83
9. The current method of typing an uppercase letter in this layout restricted me from using more uppercase letters in my passwords.	0.77
10. The current method of inserting a numeric digit in this layout restricted me from using more digits in my passwords.	0.78
11. The current method of inserting a special character in this layout restricted me from using more special characters in my passwords.	0.75
12. The current method of typing in this layout restricted me from typing a longer password.	0.78
13. I could easily type the exact letter that I wanted to type in this layout.	0.75
14. It was easy to make edits when typing in this layout.	0.75
15. It was easy to type using both hands in this layout.	0.62
16. I was able to quickly type an uppercase letter in this layout.	0.77
17. I was able to quickly insert a numeric digit in this layout.	0.82
18. I was able to quickly insert a special character in this layout.	0.81
19. I was able to quickly type passwords using this layout.	0.89

5.2.3 Initial Scale Administration – Study 1

Using the retained items, we then conducted a laboratory study for the purpose of testing the psychometric properties of the selected items. Specifically, the study was designed to not only collect responses from participants to examine their patterns, but to also examine whether participants' responses would change systematically in response to changes in the stimuli (the interface) being rated. The study was administered through the *research pool* of the department of psychology of the University of Texas at Arlington (UTA).

5.2.3.1 Participants

A total of 49 undergraduate students (28 female and 21 male) signed up and participated in our study for course credit. Written informed consent was obtained from each participant.

5.2.3.2 Material

Three layouts were used as the conditions for the study: (a) mobile phone with touchscreen keypad layout, (b) mobile phone with physical keyboard layout, and (c) computer keyboard layout. We used a within-group experimental model where each participant used all the three layouts to construct passwords.

For this study, we used a Motorola MILESTONE A853 mobile handset running Android 2.1. This handset features both a QWERTY-type touchscreen keypad and a slide-out physical keyboard. Each participant was asked to construct passwords by using both of these layouts and also a standard desktop computer keyboard.

5.2.3.3 Procedure

First, we asked each participant to construct new passwords by using the one layout for two banking websites: Chase.com and Wellsfargo.com. We wanted the participants to construct long passwords that would contain uppercase letters, digits, and special characters. To protect their security, we explicitly told the participants not to provide any of their existing passwords or any of the passwords they had previously used. For Chase.com, the participants were presented with the following scenario:

“Chase is one of the largest banks in the US and it has an ATM on campus. Imagine that you are creating an account at Chase.com for online banking. You have reached the final step of creating your new account and you need to create a strong password (a password that is long and contains uppercase and lowercase letters, digits, and special characters). Proceed to the next page to input your new password. Do not provide a password that you currently use or have previously used for any accounts. Also, do not use any confidential or personally identifiable information in your password.”

When they clicked OK, the password construction page appeared. Once they constructed the password for Chase.com, a similar scenario was presented for Wellsfargo.com.

Next, the participants were asked to type five fixed passwords. These fixed passwords were from seven to thirteen characters long and contained multiple uppercase letters, digits, and special characters (TRoub@dor!123, for example).

After a participant finished typing the fixed passwords, she was asked to evaluate the layout by using the 19-item scale. The items were randomly ordered to avoid

any ordering effects. A 5-point Likert scale (anchored by 1 = “strongly disagree”, 5 = “strongly agree”) was used to capture the participants’ responses. We note the difference between a Likert-type item and a Likert scale. A Likert-type item is a single question or statement and it falls into the category of ordinal level data. A Likert scale, on the other hand, is composed of multiple Likert-type items. The responses for the individual items are combined and then averaged to obtain a final scale score. Likert scale data are analyzed at the interval measurement scale and descriptive statistics like mean/standard deviation and statistical methods like ANOVA could be used in this regard [52].

The same process was then repeated for the second and third layouts. The order of the layouts was randomized for each participant.

Overall, each participant typed seven passwords in each layout. Out of these seven passwords, two were selected by the participant and five were given by us. The only reason for asking them to construct two of their own passwords was to ensure that they would be able to properly respond to the four items related to “restriction” (items 9-12 in Table 5.1). When administering the scale to the participants, we also modified these items slightly to emphasize new password construction. For example, item 9 was written in this way “When I was constructing a new password for the two banking websites, the current method of typing an uppercase letter in this layout restricted me from using more uppercase letters in my passwords”.

We note that we did not use deception in this study; the participants were directly asked to construct and type passwords. We also did not store any of their passwords. Given the nature of the scale and the relative lack of consequences (e.g. no embarrassment, no reason for responding dishonestly), there was no reason for hiding the true intent of the study at this stage of scale development. Similarly, participants were free to provide suggestions or concerns regarding the items and the

layouts. Upon completion of the required tasks for each condition, participants were asked to evaluate their experience by using the item list.

As each participant evaluated three layouts, we collected a data set with a total of 147 evaluations. The scores of the reverse-coded items were inversed before adding them to the data set. There were no missing data points. We used this data set to assess the reliability and the validity of our 19-item scale.

5.2.4 Reliability Analysis

We first assessed the reliability of our scale. Nunnally points out that reliability is a necessary precondition for validity [85]. There are several types of reliability estimates: inter-rater reliability, test-retest reliability, parallel-forms reliability, and internal consistency. In his landmark paper, Churchill strongly emphasizes internal consistency over the other types of reliability [25]. For a Likert scale like ours, internal consistency is the reliability estimate that is most frequently reported [44].

Internal consistency of a scale is calculated based on the covariations between different items of that scale. It measures whether multiple items that are generated to measure the same general construct produce similar scores. For example, if a participant expresses agreement with the item “It was easy to type an uppercase letter in this layout” and disagreement with the item “I felt annoyed when typing an uppercase letter in this layout”, it would indicate good internal consistency. Internal consistency can be measured statistically by calculating the Cronbach’s alpha [29].

Since Cronbach’s alpha usually increases as the covariations among items increase, a low Cronbach’s alpha value suggests that the items are possibly not measuring the same construct. Along with the Cronbach’s alpha value of the entire scale, the corrected item-total correlation values of the individual items also need to be calculated. The corrected item-total correlation value is an estimate of whether a

given item is consistent with the averaged behavior of the other items. A low corrected item-total correlation value of an item would indicate that the item should be removed, as that particular item is ultimately not discriminating participants well in regards to what the remainder of the items are measuring. Nunnally recommends removing the items with corrected item-total correlation values lower than 0.30 [86]. Once these items are removed, the Cronbach’s alpha should be recalculated to see whether a satisfactory value is achieved. However, if the value of Cronbach’s alpha is too low, a researcher should loop back to the previous step of domain specification and item generation to find out what might have gone wrong [25].

The reliability results of our data set are shown in Table 5.1. Cronbach’s alpha for the scale is 0.96, which is excellent according to the recommendation of George and Mallery [42]. This value is even arguably high, and suggests that some items could be removed and still maintain the general essence of what is being measured. We discuss this further in Section 5.4. Furthermore, all the corrected item-total correlation values were much larger than the cutoff value of 0.30, with the lowest correlation at 0.62. We therefore retained all the items at this point.

5.2.5 Construct Validity Assessment

We assessed the construct validity of our scale through a technique called the *known-groups* method [53], which involves administering the scale to conditions/groups expected to differ due to known characteristics [92]. For example, a scale to measure the construct of “fun” should show a large difference between subjects playing a video game and subjects made to wait with nothing to do. If the conditions/groups have a significant difference between their mean scores on the scale, this provides evidence for the scale’s construct validity, since this indicates that it is able to discriminate among conditions/groups that are known to be different. In other

words, this indicates that the scale effectively captures the underlying construct it is supposed to capture, which is the requirement of construct validity.

As mentioned before, we asked our participants in Study 1 to construct passwords in three different conditions. In addition to two types of mobile keypad/keyboard layouts, they were also asked to construct passwords by using a computer keyboard. The computer keyboard condition was added so that we would have a known “comfortable” condition. Constructing a strong password on a computer keyboard is easier than constructing it on a mobile keypad/keyboard due to the space constraints of the mobile device and the inconvenience of capitalizing letters and inserting digits or special characters. For example, on an iPhone, one additional click is required for each shift to and from digits, and this shift presents a different keypad view to the user. On the other hand, digits can be inserted in the same way as letters on a computer keyboard.

We compared aggregated means in the two mobile conditions to the computer condition via repeated-measure ANOVA. Mean scores for the combined mobile conditions ($M = 3.32$, $SD = .79$) were significantly lower than for the computer condition ($M = 4.39$, $SD = .68$), $F(1, 47) = 90.92$, $p < .05$. This established the construct validity of our scale.

5.2.6 Criterion Validity Assessment – Study 2

Criterion-related validity tests the relationship between a scale score and a particular outcome. For example, in the United States, SAT scores are used to determine whether a student will be successful in undergraduate studies. Here, the criterion for success for an undergraduate student may be her first-year GPA. If her SAT score correlates positively with her first-year GPA, it would indicate that her SAT

score has effectively predicted her future performance in college, thus demonstrating an evidence of the criterion-related validity of the SAT.

In our case, in order to demonstrate evidence for criterion-related validity of our scale, we selected two outcomes that are potentially related to comfort of constructing a strong password when using a particular layout:

- The length of the constructed password
- The total number of uppercase letters, digits, and special characters in the constructed password

Although there exists no empirical evidence that the comfort of constructing a strong password is related to the total number of uppercase letters, digits, and special characters, the experimental results of Haque et al. provide the primary rationale for this proposition [49]. Their results demonstrate that if users are presented with a more comfortable mobile handset interface for entering digits and some special characters, they construct passwords that contain significantly more digits and special characters [49]. As for length, we implicitly assume that the more comfort a user feels when using a particular interface, the longer her typed password would be.

In order to observe the correlation between our construct of interest and the selected outcomes, we conducted a separate study.

5.2.6.1 Participants

A total of 30 undergraduate students (17 male and 13 female) from UTA voluntarily participated in this study, and they were recruited from a course on computer literacy. The course is offered to majors from all departments and gets a diverse set of students. In exchange for their time, students were assigned extra course credits. Written informed consent was obtained from each student, and an alternative extra

credit assignment was offered to the students who were not willing to participate in our study.

5.2.6.2 Material

In this study, participants were asked to construct passwords by using one of the two layouts of our Motorola MILESTONE A853 mobile handset (see Section 5.2.3.2). Each participant was randomly assigned to one of the layouts to construct passwords. Since we collected the passwords of the participants for this study and analyzed them, we used deception in this study so that the participants would construct passwords just the way they do in real-life situations.

5.2.6.3 Procedure

We designed this study so that it appeared to the participants as if they were opening a new bank account at Chase.com. They were asked to complete a set of tasks that resembled the usual steps of creating a new online bank account. Password construction was framed as one of these multiple tasks, not as the primary task.

The participants were first presented with the following instructions:

“Chase is one of the largest banks in the US and it has an ATM on campus.

Imagine that you are creating an account at Chase.com for online banking.

Proceed to the next page to start creating your new bank account.”

When a participant clicked OK, she was asked to enter dummy values given to her on a piece of paper for the following fields: name (containing both uppercase and lowercase letters), account number, address (containing multiple special characters), phone number, and email address. These tasks ensured that the participants were familiar with the typing interface, including entering uppercase letters, digits, and special characters that would be needed for a strong password. Next the participant

was asked to answer a few questions like “Do you want overdraft protection for your new account?” and “How much daily withdrawal limit do you want?”. Finally, the participant was redirected to the password construction page where she was asked to construct a strong password for the new account. We did not enforce any requirement for length or the use of uppercase letters, digits, or special characters, though we did offer a hint for what a strong password is in our instructions:

“Please create a strong password (a password that is long and contains uppercase letter and lowercase letter, digit, and special character) for your new account. Proceed to the next page to input your new password. Do not provide a password that you currently use or have previously used for any accounts. Also, do not use any confidential or personally identifiable information in your password.”

After a participant finished all these steps, she was asked to evaluate the password construction experience on her assigned layout by using our 19-item scale. As with Study 1, the items were randomized and a five-point Likert scale was used to capture the responses.

5.2.6.4 Results

For each participant, we correlated the length of the constructed password and the total number of uppercase letters, digits, and special characters with the mean score from the scale.

Mean scale score vs. length

Mean scale score and length were strongly correlated, $r(28) = .51, p < .05$.

Mean scale score vs. total number of uppercase letters, digits, and special characters

The correlation between mean scale score and total number of uppercase letters, digits, and special characters was moderately strong, $r(28) = .41$, $p < .05$. Furthermore, we calculated mean scale scores by considering only the 12 items that are related to uppercase letter, digit, and special character (items 1-3, 5-7, 9-11, 16-18 in Table 5.1), and correlated these scores with the total numbers of uppercase letters, digits, and special characters. As expected, the correlation was stronger in this case, $r(28) = .47$, $p < .05$.

According to Cohen, a validity coefficient can be interpreted in this way: less than .1 is trivial; .1 to .3 is weak; .3 to .5 is moderate; and greater than .5 is strong [28]. Based on this guideline, our correlation coefficient values were satisfactory and evident of good criterion-related validity.

5.3 Profiling Popular Smartphone Handset Interfaces – Study 3

In order to demonstrate the practical application of our scale, we evaluated the password construction interfaces (keyboard/keypad layouts) of popular smartphone handsets through our scale. We selected three handsets: BlackBerry Curve 9300, Motorola DROID 2 A955, and iPhone 4s. The iPhone handset was selected because of its touchscreen keypad layout, while the BlackBerry and Motorola handsets were representatives of QWERTY-type keyboard and slide-out physical keyboard, respectively.

We also implemented the custom touchscreen layout proposed and designed by Haque et al. as an Android app running in a Motorola MILESTONE A853 handset [49]. It involved adding two extra on-screen rows, one containing the ten digits

and the other containing ten common special characters, in addition to the default Android touchscreen keypad.

5.3.1 Participants

A total of 21 undergraduate (15 female and 6 male) students from UTA participated in this study. As with Study 1, we recruited participants from the research pool of the department of psychology. Written informed consent was obtained from each participant.

5.3.2 Procedure

Since we did not require to collect any password constructed by the participants, we used exactly the same experimental design as Study 1 for this study. Participants were asked to type two passwords of their own and five fixed passwords by using each of the four layouts. After typing the passwords by using one layout, they evaluated that particular layout by using our scale.

5.3.3 Results

For each layout, we calculated the mean scale score. The iPhone 4s touchscreen keypad layout was rated the most comfortable (mean = 4.19 out of 5), while BlackBerry's layout was considered the least comfortable (mean = 2.78 out of 5). The Motorola layout received a moderate score (mean = 3.32 out of 5). The custom layout of Haque et al. obtained a slightly lower score (mean = 4.13 out of 5) than that of iPhone 4s.

We note that these findings should be interpreted with caution. We discuss this in detail in Section 5.5.2.

5.4 Factor Analysis

Factor analysis is a statistical procedure that examines the correlations or covariances among items to discover clusters of related items. In psychometrics, factor analysis is often used to identify the underlying *subconstructs* that might reside in the construct of interest. These subconstructs are also referred to as *factors*, *components*, or *dimensions*. For example, in his classic paper, Spearman uses factor analysis to posit a two-factor theory for measuring human intelligence: the general intelligence factor and the specific intelligence factor [110].

Factor analysis comprises two different perspectives: exploratory factor analytic approaches and confirmatory factor analysis. Exploratory factor analysis is used when a researcher is uncertain about the theoretical conceptualization of her construct of interest. It provides a quick way to explore the underlying factors of the construct, thus providing an opportunity to refine the theory at an early stage of scale development. Confirmatory factor analysis (CFA), on the other hand, is used when the researcher has a more specific theory about the conceptualization of the construct of interest. Based on this theory, the researcher builds a model and gathers data to examine whether the data fits the hypothesized model.

Factor analyses can provide meaningful information regarding the overarching structure of the data, and can provide guidance on how best to aggregate the data after the factor. There are a number of ways to extract factors, including principal component analysis (PCA), principal axis factoring (PAF), maximum likelihood, and more, but PCA and PAF are most frequently used. Factor rotation is an important consideration during a factor analysis. By maximizing high item loadings and minimizing low item loadings, rotation helps to produce a more interpretable factor analysis solution. There are several rotation techniques, varimax rotation is the one that is used most commonly.

We conducted a PCA with a varimax rotation (eigenvalues greater than 1) on our data set for Study 1 [61], and it was found that items loaded on one general component. The one component accounted for 62% of the variance, and item loadings ranged from 0.61 to 0.91. PCA tends to however identify one factor, and does not allow for examination of more complicated models as is possible in CFA.

Given the strength of relations obtained across items (demonstrated via both the Cronbach’s alpha and the PCA), we decided that there was undue redundancy in the items and decided to cut unnecessary items. Upon careful examination of the current items and their respective relations, we were interested to examine the extent to which a higher order factor (comfort), and four corresponding second level factors (uppercase letter, numeric digit, special character, and general typing) would fit the data based on the eight retained items.

Our hypothesis was based on the observation that issues like ease of edit (“It was easy to make edits when typing in this layout”), ability to type by using both hands (“It was easy to type using both hands in this layout”), and ability to type the exact letter that the user wants to type (“I could easily type the exact letter that I wanted to type in this layout”), in turn, result in quick and easy typing of passwords. Items 13, 14, and 15, therefore, essentially capture the quickness and easiness of general typing when using a specific layout. Furthermore, the items related to frustration and restriction actually capture the cognitive and emotional hindrances of a user when constructing a strong password by using a particular layout. Intuitively, these hindrances should prevent users from typing quickly and influence their perceptions regarding the ease of using the layout.

We therefore focused exclusively on the quickness and easiness related items and posited the following four-factor theory regarding the comfort of constructing a strong password when using a particular layout:

Factor: Uppercase letter

1. It was easy to type an uppercase letter in this layout.
2. I was able to quickly type an uppercase letter in this layout.

Factor: Numeric digit

1. It was easy to insert a numeric digit in this layout.
2. I was able to quickly insert a numeric digit in this layout.

Factor: Special character

1. It was easy to insert a special character in this layout.
2. I was able to quickly insert a special character in this layout.

Factor: General typing

1. It was easy, overall, to type passwords using this layout.
2. I was able to quickly type passwords using this layout.

We used IBM SPSS Amos (version 21) to conduct CFA and evaluate the fit of our confirmatory model. The default settings (i.e. maximum likelihood) were used, with the raw data of Study 1 supplied as an input. It was found that our proposition was supported, the data fit the overarching model well ($\chi^2(16) = 24.952$, $p = .07$, $RMSEA = .06$, $PCLOSE = .31$). Contrary to other statistical models, the null hypothesis is that the model fits the data well. Thus, a chi square value that does not reach statistical significance is considered indicative of good fit. Because of the extremely conservative nature of this particular statistic (i.e. rarely do arguably good fitting models meet this criteria), RMSEA and PCLOSE statistics are also typically reported. A small RMSEA value is an indicator of good fit as a value of 0.08 or less is often considered acceptable [17]. PCLOSE is a test of statistical significance for

RMSEA, with the assumption that the RMSEA= .05 (i.e. close fit). A statistically significant difference again means that the theoretical model is significantly different from the actual relationships among variables (which is not in our case, hence a good fit). Thus, the statistical results demonstrate that our model is likely a good fit for our data.

These results suggest that there appears to be four highly related factors in the scale that collectively comprise our representation of user comfort. In turn, data can be averaged to the level of the scale for most purposes (or in the case of missing data, data should be averaged at the level of factor, and then the factors averaged for representative individual indicators). Similarly, if variations in comfort based on these factors need to be examined (e.g. “Is this particular layout more comfortable for typing uppercase letters?”), then the scale can effectively do so by specifically examining those specific factor values.

5.5 Discussion

This is the first study to date that we are aware of which specifically applies psychometric principles to develop and test a scale designed to measure how well suited keyboard or keypad layouts are in the context of password construction. We have utilized numerous frameworks and conceptualizations, and extensively tested the scale in various ways to create the most accurate, useful scale possible. From extensive content validation efforts, to examination of construct validity and analysis of factor structure, to prediction of important meaningful criteria, this scale has demonstrated very promising initial evidence.

In the subsequent sections, we first discuss the ecological validity of our study and highlight the limitations of our work. Next, we discuss several issues related to scale development.

5.5.1 Ecological validity

As mentioned before, Study 1 and Study 3 did not involve deception, and we did not try to hide the motive of our study from the participants for these two studies. This was in accordance with the experimental methodology for scale development studies, where users are first explicitly subjected to a certain task and later asked to evaluate the experience by using the scale. For our case, the task was to type a few passwords (five fixed, two of the users' own) by using different layouts. The experience of constructing a password was more important here, rather than the password itself. When the users were constructing their own passwords, we involved a simple role-play scenario, since prior work has shown that it is more effective than a survey scenario in motivating the users to construct passwords more seriously [69].

For Study 2, we collected passwords constructed by the participants. Ecological validity therefore was an important consideration for this study. The results of a recent work of Fahl et al. on the ecological validity of password study reveal that passwords collected during user studies closely resemble users' actual passwords [35]. We tried our best to frame Study 2 as an experiment that asks the users to perform a real-life online task, namely creating a new online bank account by using a mobile phone handset. Password construction was one of a series of steps for completing the primary task (i.e. creating the new account), just as it would be in real life. The word "password" was not used anywhere in the informed consent document. A debriefing session was arranged at the end of the study where the deception was revealed and the participants were provided with the opportunity to withdraw their consent to participate in the study. None of the participants decided to do so and we could use all of their passwords to test the criterion-related validity of our scale.

We note, however, that our participants were not required to return on a second day to re-enter their passwords, and as such, we were not completely able to emulate the real-life password construction behavior of users.

5.5.2 Limitations

For this work, we quantified password strength in terms of entropy, according to the recommendation of password researchers (see Section 5.2.1). We do not overlook the findings of Weir et al. or Kelley et al., which demonstrate that entropy is not the most appropriate measure of password strength [118, 65]. However, since our developed scale focuses on measuring the comfort of constructing a strong password when using a particular layout, we believe that entropy is a better approximation of password strength here because it effectively captures the layout-related aspects of a strong password. Alternative measures such as guessability are more dependent on the exact password choices of users and do not clearly capture aspects related to keyboard layout, such as the use of special characters. This approximation is consistent with Haque et al. [49], a related work on password strength and keypad/keyboard layout.

For all three studies, we recruited participants from university students, who may vary considerably from other populations in their smartphone usage behavior. We plan to test our scale by using a more diverse population group in future. Ultimately, scale development is a never-ending process in which developers continually strive to understand the intricacies of the conventions in regards to any meaningful variations (e.g. does my scale predict other meaningful criteria, does it behave differently in other contexts or for other sample compositions, etc.). However, in general, this scale has demonstrated solid initial evidence of its efficacy.

Our results of Study 3 should be interpreted carefully, particularly by considering the fact that we did not control for participants' previous familiarity with the

interfaces. For example, if most of the participants were iPhone users, their familiarity with the iPhone layout would probably bias them towards that layout. We conducted Study 3 for demonstrating a practical application of our developed scale, not for a definitive comparison among the interfaces.

5.5.3 Aggregation and application

Our shortened item list has four factors, each of which contains two items (see Section 5.4). Since each of the underlying factors contains the same number of items, none of the factors is underestimated or overestimated when individual item scores are combined and averaged to form a final composite score. Subsequently, depending on the intended application (examination of individual level issues with comfort, or identification of problem areas with the layout), the scale score could also be computed in terms of each factor. As a result, the scale could be used to answer more specific questions like “Which layout is more comfortable for inserting a numeric digit when constructing a strong password?” or “Which layout is more comfortable for general password typing?”. This provides an additional motivation for us to conduct further experiments with this shortened scale.

5.5.4 Norm development

After a sound scale (reliable and reasonably valid) has been developed, depending on the intended application of the scale, the researcher should continue to conduct further experiments. If the purpose of the scale is to compare different interfaces with respect to the construct of interest, then administering the scale to different users and profiling the interfaces based on scale scores should be sufficient. Our scale can currently be used in this way.

On the other hand, if the purpose of a user comfort scale is to answer the question of whether users are sufficiently comfortable with a particular user interface, then the researcher should also develop norms for her new scale. Developing norms involves setting up standard scores for a scale. Ideally, for a 5-point Likert-type scale, mean scale scores of 3 and 4 should imply neutral and positive attitudes, respectively. However, this might not be always true. For example, a mean score of 4 might represent the highest (or lowest) score ever achieved on that particular scale. To this end, the researcher should specify the benchmark scores for her new scale. This can be done by administering the scale over a large number of users to obtain a distribution of scores and subsequently characterizing the distribution by various statistical features such as mean and standard deviation. A detailed description about the norm development procedure can be found in [43].

We believe that this norm development technique could be used to specify a standard score that would represent “sufficient user comfort” in the context of a specific security system user interface. This would be helpful to precisely find out whether users are sufficiently comfortable with a particular security system user interface, which, according to the working definition of usable security in the seminal paper of Whitten and Tygar [119], is an important consideration for measuring the usability of that security system.

5.5.5 Revalidation study

We note that we did not prune any items during the reliability assessment stage because all of the items had a satisfactory corrected item-total correlation value and the Cronbach’s alpha value of the overall scale was high (see Section 5.2.4). If items need to be pruned at this stage, a revalidation study is recommended to be conducted with the shortened scale. This involves administering the shortened scale to a new

sample which is independent to the previous sample and assessing the reliability of the shortened scale.

For our scale, however, we needed to assess the criterion-related validity by using a separate study that involved deception and collection of participants' passwords. This provided us an opportunity to reassess the reliability of our scale by using a different sample. We calculated the Cronbach's alpha for this new data set. As before (0.96), the value was high enough (0.93). This provided further evidence for the reliability of our scale.

CHAPTER 6

Novel Techniques for Memorizing System Assigned Random Passwords

First we describe the constraints and objectives that define the design requirements of our system.

6.1 Constraints and Objectives

In a system-assigned random password scheme, the system generates a sequence of random characters to be used as a password. The total number of characters depends on the desired entropy of the random password.

6.1.1 Entropy

We calculate entropy by using the formula entropy $H = L \log_2 N$, an approximation of plain Shannon entropy [105], where L is the length of the password and N is the size of the alphabet. Although a larger number of N could be obtained by including digits and/or uppercase letters, we consider lowercase letters only since this would further allow us to extend our study for mobile devices in the future, as previous works have demonstrated the inconvenience of inserting digits and uppercase letters when constructing a password by using a mobile device [50, 48].

As we fix the size of the domain for N to 26 (lowercase letters only), we obtain the value of L by selecting our desired entropy level. A study on security policies suggests that 20 bits of entropy (with lockout rules) should suffice for protection against online brute-force attacks [37]. For five lowercase letters ($L = 5$), the entropy would be $5 \log_2 26$ or approximately 23.5 bits. We choose a more conservative approach and

select six lowercase letters ($L = 6$), which yields an entropy value of approximately 28 bits, well above the threshold value of 20 bits. Prior studies on system-assigned random passwords also selected this desired level of entropy [123]. Furthermore, 28 bits of entropy provides password-level security [9].

In our system, we propose to randomly assign the users a password with six lowercase letters ('rlspxh' for example) and then introduce our novel training methods to make them drill with the password and memorize it effectively. Since password construction itself is a part of the registration process on a website, our system essentially adds another step to the process by introducing the password training/drilling method. Thus, time is an important design consideration for our system to avoid unreasonably longer registration time.

6.1.2 Time

We note that unlike logging in, registering to a website is a one-time activity. Our training methods are used during the registration period only, not during logging in. Users already need to perform a variety of tasks during the registration period. For example, some sites ask the users to provide credit card information during registration, before allowing them to try their service out. We therefore believe that our training activity would not disrupt the users much since they are already accustomed to performing a wide range of tasks during the registration period.

At the same time, the training activity should not be long enough to make them annoyed. Thus, we select three minutes as the maximum duration of the training activity, which we believe to be a reasonable one¹. We note that the activities will

¹All of the participants in our pilot study (n=8) expressed their satisfaction with this time duration and reported that they would be willing to spend this amount of time for memorizing a password for one of their important accounts.

be designed in an interesting way to avoid boredom (e.g., watching a video clip). We describe this in detail in Section 6.3.

6.1.3 Automation

Along with many other security researchers, we acknowledge that: “Security is only as good as its weakest link, and people are the weakest link in the chain” [100]. Thus, we decide to not involve the users in doing something on their own to facilitate the memorization process (such as imagining some pictures of their own preferences or creating their own stories), rather we plan to generate the training contents automatically based on the assigned random passwords. This makes our study design essentially different from existing works on memorization techniques. We discuss this in next section.

6.2 The Memory Techniques

We extensively reviewed the literature on memory and memorization techniques to identify the most potential techniques for memorizing system-assigned random passwords. Specifically, the requirements defined in the previous section helped us to narrow down the list of potential techniques. A good number of techniques focus on memorizing digits, such as *phonetic system* and *phonetic recoding* [32]. We excluded them since we exclusively focus on memorizing letters only (§6.1.1).

Other popular methods like *peg* or *hook* system require users to learn a series of memory pegs first [88], which is not compatible with our time and automation requirements (§6.1.2 and §6.1.3). All such techniques that require learning something in advance were excluded. Some techniques were excluded since they are proven to be inefficient, such as *imagery* method [97].

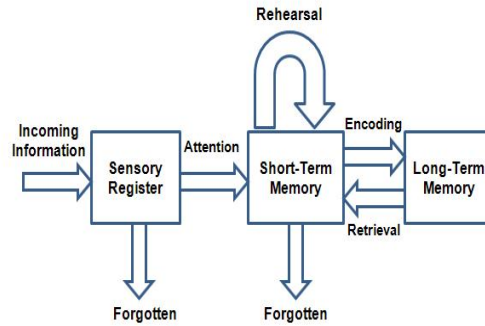


Figure 6.1: Atkinson-Shiffrin Memory Model

After this filtering step, we pinpointed two potential techniques that are compatible with our requirements: the ancient *method of loci* (also known as the *memory palace* method), and the *link* or *story* method. We now discuss the scientific implication of each of these techniques.

6.2.1 The Science Behind Proposed Techniques

First we present a general theoretic framework of human memory, as proposed by Atkinson and Shiffrin in their landmark paper [6]. This framework identifies three basic structural components of human memory: the sensory register, the short-term memory, and the long-term memory (Figure 6.1 illustrates the framework). According to their theory, any new information first enters the sensory registers, where it retains for a very brief period of time, before eventually getting decayed and lost. The short-term memory, which also acts as the human working memory, receives selected inputs from the sensory registers. The information retains in short-term memory for about 30 seconds. However, a control process called rehearsal can retain a limited amount of information in short-term memory for a much longer period of time. The final component, the long-term memory, acts as a fairly permanent repository of in-

formation for humans. Information transfer from short-term memory to long-term memory depends upon further processing and encoding. In this regard, if the information could be associated with something meaningful, the encoding process would be more elaborative. The memorization techniques assist for elaborative encoding by providing meaningful cues.

The paramount importance of cues has been highlighted in another landmark paper of cognitive psychology, where Tulving introduces the concept of cue-dependent forgetting [116]. He proposes that memories are not lost or forgotten; forgetting something essentially means that the necessary cues needed to retrieve them are unavailable. Thus, cues are considered to be fundamental source of information during a retrieval process.

6.2.1.1 The Method of Loci or the Memory Palace Method

The *method of loci* (also known as the *memory palace* or *mind palace* technique) is one of the oldest mnemonic techniques that have been used extensively to facilitate memory recall [125]. The method has been extensively used in other domains for assisting memory-impaired persons [95], investigating memory plasticity for children as well as adults [7], and examining the memory decline effect of adults [126]. A design of a graphical password scheme has been proposed which was inspired by the *method of loci* [75].

For using this technique, one first needs to identify a few landmarks (also known as *loci*) in some familiar place, such as her home or office building. When requiring to remember a set of items, she mentally walks through these landmarks and associates each item with a particular landmark. This association can be made by forming a vivid visual image of the item and placing it on the landmark. For recalling the items

later, she re-imagines walking through the landmarks, and retrieving the items in order.

The utilization of this method has been exemplified in TV series “Sherlock”, a BBC-produced crime drama series that portrays a contemporary adaptation of Sir Arthur Conan Doyle’s famous Sherlock Holmes detective stories [33]. During an episode in Season 3, titled as “His Last Vow” [80], Sherlock can be seen using his *mind palace* for discovering the best path to survival².

This technique leverages spatial memory, the part of memory that helps recording information about one’s surrounding environment and the associated spatial orientation [18]. More specifically, using this technique activates portions of the brain that are concerned with spatial awareness, such as the medial parietal cortex and the retrosplenial cortex [77, 90]. Both of these regions play an important role in enhancing spatial learning abilities. The *method of loci* technique also activates the right posterior hippocampus [77, 90], an important component of the brain which assists in finding one’s way around an environment and remembering the associated events which occur within it [18].

Research results have revealed that memory champions do not have extraordinary brains; rather they use their normal spatial abilities to great effect [94, 77]. This makes us think that regular users can also take advantage of these abilities, if they have guidance to facilitate quick adoption of the *method of loci* via our training interfaces.

²Although Sherlock’s *mind palace* does not resemble the typical type of storage place for the *method of loci*, it gives an idea of organizing information in a certain way to facilitate the information retrieval from memory.

6.2.1.2 The Link or Story Method

Link method is another mnemonic technique which is simple to use and requires no set of materials to learn in advance [76]. The basic concept of the *link* or *story* method has been applied in a graphical password scheme. In that scheme, the password has been proposed to be a sequence of images selected by a user from a pool of larger images to make a story [30].

Applying the *link method* involves converting each item to be remembered into a pictorial representation and creating a link or association between each successive pair of representations in a vivid way. This, in turn, creates a chain of interacting images (items) where the first image acts as the cue to recall the second one, the second image acts as the cue to recall the third one, and so on.

Research results have shown that such linking of images of items together improves recall performance relative to forming images of each item individually [81]. More interestingly, subjects that conducted the latter task performed no better than uninstructed control subjects in recalling the items. This highlights that imagery instructions work best when they are linked together to form interactive images. Since this linking of images ultimately lead to a cohesive story, the technique is also known as *story method*.

A separate study assessed the memorability of four memorization techniques, including the *link method* and the *method of loci* [97]. The results demonstrated that the *method of loci*, along with the *peg method* outperformed the other techniques. The *link method* performed better than the *imagery* technique, as well as the control condition. These results inspired us to select our techniques. However, as mentioned before, the *peg method* was not considered due to its incompatibility with our time and automation requirements (§6.1.2 and §6.1.3).

6.3 System Design

Our system has two main parts: registration and login. During the registration period, we first assign a system-generated random textual password consisting of six lowercase letters (§6.1.1) to a user. Next, we dynamically generate a video clip based on the letters of the assigned password and show the clip to the user. The video clip is a quick adoption of the *method of loci* or the *link method*, which assists the user in memorizing the random password. Once the clip ends, the user is asked to type the password to complete the registration phase.

During the login period, users just need to recall and type the correct password to authenticate successfully, just like the way they do in any recall-based textual password scheme. This simple and straightforward recall-based authentication scheme has two major advantages. First, it avoids a longer login time, which is a major usability issue for any recognition-based authentication scheme [4]. The second major advantage is that it makes the deployment procedure very simple, nothing extra needs to be done during the login period. We discuss this in detail in Section 6.6.

Our work also differs from the related works in cognitive psychology in a major way [81, 97, 72]. We do not require the users to do anything on their own, such as imagining some pictures of their own preferences or creating their own stories, for facilitating the memorization process. This essentially removes the susceptibility associated with any kind of poor user action or selection.

We now give a detailed description of our video clips for the *method of loci* and the *link method*.

6.3.1 Method of Loci

Although mentally navigating a familiar environment seems to be more effective for carrying out the *method of loci*, research results have demonstrated that virtual

environments are as effective as familiar environments in this regard [72]. We therefore created a virtual environment for implementing the *method of loci*.

The virtual environment was modeled after a real world apartment consisting of a living room, a kitchen, a dining room, a bedroom, and a washroom. These five rooms, along with a mail box at the entrance of the apartment, served as the six loci. The layout of the apartment was fixed: mail box at the entrance, followed by living room, kitchen, dining room, bedroom, and washroom, in that order. We also selected 26 distinct objects to pictorially represent 26 alphabets (apple for ‘a’, ball for ‘b’, and so on). Depending on the generated random password, six out these 26 objects would appear at the six loci of our virtual apartment.

For example, if the randomly generated password is ‘pcgbhr’, then the video clip would show a pencil above the mail box at the entrance, a cat above a sofa at the living room, a guitar beside the sink in the kitchen, a ball above the dining table, a helicopter above the bed in the bedroom, and a rocket inside the washroom cabinet.

At first, the video clip would show the layout of the virtual apartment, without placing any objects in the specified loci. This would familiarize a viewer with the apartment and show her the order of different loci in the layout, starting from the mail box at the entrance and ending at the washroom cabinet. Once this is done, the camera would return to the mail box at the entrance and navigation would re-start. This time the objects would be appearing in the specified loci. This process would be repeated for one more time. Thus, there would be three navigations of the entire apartment (one without and two with the objects).

For helping a viewer to better recognize the objects, the camera would zoom in whenever the objects would be shown. The navigation would pause for a few seconds and the object name would appear on top, with the first letter highlighted. The entire duration of the video clip, including these pauses, would be 160 seconds.

6.3.2 Link Method

As with the training interface for the *method of loci*, this method also involves showing a video clip to the users. However, no spatial orientation was used for this method. Instead, we planned to show a small story which would be associated with the six letters of the randomly assigned password.

For this method also, we used 26 distinct objects to represent 26 different letters. In addition, for the sake of creating an interesting story, we used 26 different animals for representing the last two letters of the password.

As before, we assume that the randomly generated password is ‘pcgbhr’. The first four letters would be represented by four objects in our story, while the last two letters would correspond to two animals. The story would begin with a magician showing his tricks in front of a small audience. The magician would start with a pencil (pencil for ‘p’, the first letter of the password) and magically transform it into a cat (cat for ‘c’, the second letter). We note that the pencil would act as a cue for remembering the cat here, since the former would be transformed into the latter one.

In the subsequent scene, the magician would place the cat above a table and ask for a volunteer from the audience. A girl would volunteer and the magician would transform his power to the girl by uttering his magic words. The girl would then magically transform the cat into a guitar (guitar for ‘g’, the third letter). At this point, the magician would ask for a volunteer boy to join him at the stage. Once the boy arrives, the magician would magically lift the guitar from the table and use it to hit a magic cloud on the roof. As a result, there would be magic rain of balls (ball for ‘b’, the fourth letter).

One of the balls would hit the boy, resulting him to be magically teleported to a jungle, where he would encounter a horse (horse for ‘h’, the fifth letter). Seeing the horse, the boy would get afraid and scream for help. The magician would then send

a magical rabbit (rabit for ‘r’, the sixth letter) to save the boy from the horse. At the very last scene, the chain of cues would be shown to the viewer: pencil \rightarrow cat \rightarrow guitar \rightarrow ball \rightarrow horse \rightarrow rabbit.

The entire video clip would be accompanied with appropriate captions, describing the sequence of the incidents. As before, the camera would zoom in and the navigation would pause whenever the objects would be shown. The entire duration of the clip would be 180 seconds for this method.

We note that the basic organization of the story would remain the same, regardless of the assigned password. However, the objects (e.g., apple instead of pencil) and the animals (e.g., tiger instead of horse) would change, depending on the letters of the assigned password.

6.3.3 Pilot Study

We conducted a pilot study with eight participants to primarily evaluate the effectiveness of our video clips. Their feedbacks helped us to finalize certain design issues such as the amount of light inside the virtual apartment, the duration of zooming and pausing, the size and placement of the captions/texts etc. For both of the methods, seven out of the eight participants were able to recall the memorized password after a week.

6.3.4 Development Platform and Tools

To develop our virtual apartment model and use it in custom way, we used two softwares: Max3D and Unity3D. First we developed the apartment model by using Max3D. The 3D objects representing the password letters were also modeled by using Max3D. Next we imported the model file into Unity3D game engine. We used the

Unity3D game engine to implement the camera navigation, create wall textures, and set up point light sources to make the objects clearly visible inside the apartment.

For showing the story of the magician, we basically used a series of image frames. These frames were designed by using Adobe Photoshop CS5. Later we merged all these frames by using Unity3D.

For both of the video clips, we dynamically placed different objects in different locations/frames, depending on the letters of the randomly assigned password. This logic was implemented by writing scripts for each functionalities in C Sharp.

6.4 User Study

In this section, we present the design of our user study, where we used a within-subjects design consisting of three study conditions. A within-subjects design has two major advantages: it controls for individual differences, while permitting the use of statistically stronger hypothesis tests. The study procedures were approved by UTA Institutional Review Board (IRB) for human subjects research.

6.4.1 Participants, Apparatus and Environment

For this experiment, we recruited 52 students (34 women, 18 men) through our university's Psychology Research Pool. Participants came from diverse backgrounds, including majors from Psychology, Business, Nursing, Biology, and Music. The age of the participants varied between 18 to 31 with a mean age of 20. Each participant was compensated with course credit for participation and was aware that her performance or feedback in this study would not affect the amount of compensation.

For assessing the effectiveness of our training methods and performing a better comparison, we administered a control condition where participants were asked to memorize their system-assigned random textual password in any way they preferred.

The password entropy for this condition was also 28 bits and the time limit for memorizing the password was 180 seconds (consistent with the other two methods).

The lab studies were conducted with one participant at a time to allow the researchers to observe the users' interactions with the system. We created three realistic and distinct websites using the images and layouts from familiar commercial sites, where each of them was equipped with one of our three password schemes: Control, Loci, and Link.

6.4.2 Procedure

Our experiment consisted of two sessions, each lasting around 30 minutes. To test users' memorization of the assigned passwords, the second session took place one week after the first one. This one-week delay is larger than the maximum average interval for a user between subsequent logins to any of her important accounts [54] and is also a common interval used in authentication studies (e.g., [82, 123, 34, 3, 4]).

6.4.2.1 Session 1

In the first session, the participants were given an overview of our study after signing a consent form. Then they performed registration for each of the three sites, each outfitted with a distinct scheme. For the *method of loci* and the *link method*, the registration process was exactly the same as described in Section 6.3. For control condition, users were first assigned a system-generated random password. Next they were given 180 seconds to memorize the password in any way they preferred.

After registering with each scheme, participants performed a practice login with that scheme. We did not collect data for these practice trials. During registration, the sites were shown to the participants in random order to compensate for ordering effect. They were asked to not record (e.g., write down) the assigned passwords.

Before leaving, we reminded each participant to show up for the second part of the study after a week.

6.4.2.2 Session 2

After a week, when the participants returned for the second session, they were asked to log into each of the three sites using the assigned passwords. The sites were shown to the participants in random order. They were allowed to make a maximum of three attempts for a successful login. After they had finished, we conducted an anonymous paper-based survey. Participants were then compensated and thanked for their time.

6.4.3 Ecological Validity

Our participants came from diverse majors. They were young and educated, which represents a large number of frequent Web users, but may not generalize to the entire population. We were only able to gather data from 44 participants (eight participants did not show up for the second session) since the study was performed in a lab setting. However, lab studies have been preferred to examine brain-powered memorability of passwords [35]. Moreover, since lab study is conducted in a controlled experimental setting, it helps to establish performance bounds and determine whether field tests are worthwhile in future research. We believe that 44 provides a suitable sample size for a lab study as compared to the prior studies on password memorability [115, 22, 23, 120, 3, 4].

6.5 Results

In this section, we discuss the results of our user study described in Section 6.4. We evaluated the usability of our study conditions via all metrics suggested in the

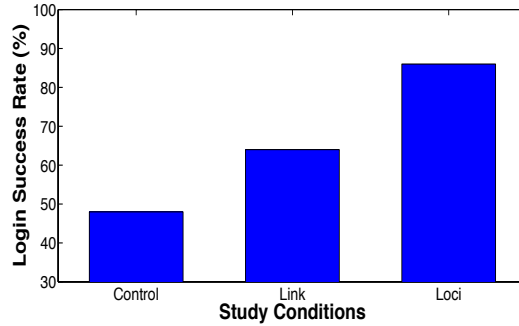


Figure 6.2: Login success rates for the study conditions [N = 44]

literature [102]: memorability (§6.5.2), registration time (§6.5.3), number of login attempts (§6.5.4), login time (§6.5.5), and user feedback (§6.5.2, §6.5.3). Since eight of the participants did not show up for the second session, we excluded their data and present our results for 44 participants.

6.5.1 Significance Tests

We use statistical tests to analyze our results, where the results comparing two conditions are considered to be significantly different when we find $p < 0.05$. When comparing two study conditions where the variable is at least ordinal, we use either Wilcoxon signed-rank test (for matched pairs of subjects) or Wilcoxon-Mann-Whitney test (for unpaired results). Wilcoxon tests are similar to t-tests, but make no assumption about the distributions of the compared samples, which is appropriate to the datasets in our conditions. Whether or not a participant successfully authenticated is a binary measure, and so we use a McNemar’s test for matched pairs of subjects and a chi-squared test for unpaired results to compare login success rates between two conditions.

6.5.2 Memorability

Our results show that out of 44 participants in the study, 38 participants (86.4%) succeeded to log in using Loci method, while 28 participants (63.6%), and 21 participants (47.7%), respectively, logged in successfully with the Link method and Control condition (see Figure 6.2). Whether or not a participant successfully authenticated is a binary measure, so we compare login success rates between conditions using McNemar’s test. Our analysis shows that the login success rates for Loci method, $\chi^2(1, N = 44) = 13.47, p < 0.01$, and Link method, $\chi^2(1, N = 44) = 2.77, p < 0.05$, were significantly higher than that for the Control condition. We also found that Loci method had a significantly higher login success rate in comparison to the Link method, $\chi^2(1, N = 44) = 5.79, p < 0.05$.

For Control condition, participants were given the opportunity to memorize the password in their own way. We observed that most of them either followed the basic repetition technique, or tried to create a mnemonic based on the assigned password. On the other hand, in Loci and Link methods, users were shown a set of objects in a specific context where each object corresponded to one letter of the system-assigned random password (e.g., “apple” corresponded to the letter “a”). The Loci method also assisted to memorize the sequence of letters by leveraging the spatial memory. The Link method achieved the same goal by creating a chain of cues. All these additional features facilitated the processing and encoding of the authentication information to store them in long-term memory (Figure 6.1). Thus, the memorability (i.e., login success rates) for Loci and Link methods were significantly higher than that for the Control condition.

The Loci method also had a significantly higher recall rate than the Link method. This result is consistent with the experimental results of a prior study

which was conducted to assess the effectiveness of these methods in the context of memorizing a list of words [97].

At the end of second session, we asked participants to answer a 5-point Likert-scale question (1: *strong disagreement*, 5: *strong agreement*) regarding the efficacy of Loci and Link methods in providing satisfactory memorability (e.g., “The passwords were easier to remember because of watching the clip.”). The results for Wilcoxon signed-rank test (appropriate for matched pairs of subjects) show that user feedbacks were significantly better for Loci method (Median: 4, Mode: 5) as compared to the Link method (Median: 3, Mode: 2) in terms of the effectiveness of a scheme to ease memorization ($V = 398$, $p < 0.05$)³.

We also asked participants if they would require writing down the authentication secret for memorability, where we reversed this question (e.g., “I would require writing down the passwords, even after seeing the clip”) to avoid bias; the scores were reversed before calculating the modes and medians. So, a higher score always indicates a more positive result for a scheme. We found a significantly better user feedback for Loci method (Median: 4, Mode: 5) comparing to the Link method (Median: 4, Mode: 4) in terms of the requirement to write down passwords for memorability ($V = 182$, $p < 0.05$).

³Since Likert scale data are ordinal, it is most appropriate to calculate mode and median for Likert-scale responses [96].

Table 6.1: Number of Attempts for Successful Logins [SD: Standard Deviation]

Study Conditions	Mean	Median	SD
Control	1.1	1	0.3
Link	1.1	1	0.5
Loci	1.1	1	0.2

6.5.3 Registration Time

The registration time was constant for Loci (160 seconds) and Link methods (180 seconds). For performing a fair comparison, we allowed the participants to spend the same amount of time (180 seconds) in Control condition to memorize their passwords.

At the end of second session, we asked for the perception of participants on the registration time of Loci and Link methods through a 5-point Likert scale question (e.g., “The time spent for learning the password was worth it”). We found a significantly better user feedback for Loci method (Median: 4, Mode: 5) comparing to the Link method (Median: 3, Mode: 2) in terms of registration time, i.e., the time for learning a system-assigned password ($V = 224$, $p < 0.05$).

6.5.4 Number of Attempts

In this paper, *number of attempts* and *login time* respectively refer to the required attempts and time for successful logins only, unless otherwise specified. We do not get matched pairs of subjects while comparing two schemes in terms of login time or number of attempts for successful logins, since some participants who logged in successfully for one scheme failed in the other scheme. So, we use a Wilcoxon-Mann-Whitney test (appropriate for unpaired results) to evaluate two schemes in terms of number of attempts and the time for successful logins.

The mean number of attempts for a successful login was less than two for each of the three study conditions, while the median was one in each case (see Table 6.1). The results for Wilcoxon-Mann-Whitney tests found no significant difference between any pair of study conditions in terms of the number of attempts for a successful login: Link-Control ($W = 312$, $p = 0.88$), Loci-Control ($W = 434$, $p = 0.58$), and Loci-Link ($W = 544$, $p = 0.72$).

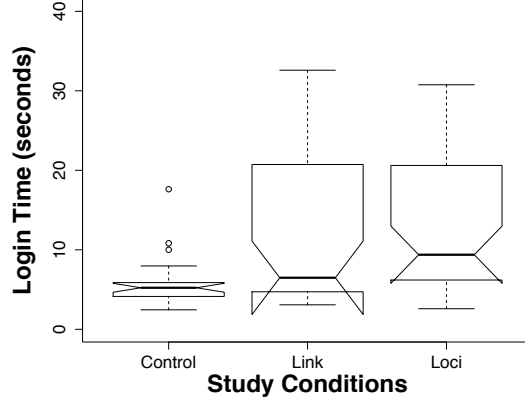


Figure 6.3: Login time for the study conditions

6.5.5 Login Time

We illustrate our results for login time in Figure 6.3. We found that the median login time for the Control, Link, and Loci conditions were 5 seconds, 6 seconds, and 9 seconds, respectively. The results for Wilcoxon-Mann-Whitney tests show that the login time for Control condition was significantly less than that for Loci method ($W = 180.5$, $p < 0.05$) and Link method ($W = 218.5$, $p < 0.05$). We did not find a significant difference in login time between Loci and Link conditions ($W = 695.5$, $p = 0.26$).

For Loci and Link methods, participants were required to recall the series of events that took place in the video clips. They were also required to recall the objects representing the password letters. Thus, the login times were slightly longer. However, compared to the login time of other recognition-based schemes [4], these values are nominal.

6.6 System Analysis

In this section, we present a detailed analysis of our system with respect to the usability-deployability-security (“UDS”) evaluation framework, as proposed by

Bonneau et al. in their work [11]. This framework suggests and defines 25 usability, security, and deployability benefits that are expected from an ideal Web authentication scheme. While it is not possible for a single scheme to provide all these benefits, their framework is useful for comparative evaluation among multiple authentication schemes.

Two things are worth mentioning before we present our system analysis in light of the usability-deployability-security evaluation framework. First, The framework exclusively focuses on Web passwords. We note that textual passwords are also largely used for authentication purposes on personal laptops and organizational desktop machines. Our scheme fits well for these types of authentication purposes as well.

Secondly, the framework acts as a benchmark for any new proposed authentication scheme that tries to replace textual passwords. Unlike numerous attempts made in prior studies [9], our aim is not to replace textual passwords, rather make it stronger. To achieve this goal, we propose a novel approach to accommodate the desired properties of an ideal authentication scheme, through leveraging the findings from cognitive psychology and existing password literature.

As a result, we do not expect our scheme to offer all the benefits that have been suggested in the framework. However, considering the comprehensiveness of the framework, we find it useful to discuss our system by relating it to the benefits that have been defined in this framework.

The usability-deployability-security evaluation framework suggests and defines eight usability, six deployability, and eleven security benefits.

Usability Benefits

1. *Memorywise-Effortless*: A scheme is said to be Memorywise-Effortless if users do not require remembering any authentication secret at all for using the scheme. Since our scheme resembles the traditional recall-based authentication mechanism, it does not offer this benefit. However, our experimental results showed that the *method of loci* achieved a high memorability rate (86%). Moreover, users expressed a high degree of satisfaction regarding the efficacy of the *method of loci* in providing good memorability (§6.5.2).

2. *Scalable-for-Users*: A scheme provides this benefit if using the scheme for multiple accounts does not create any additional burden on the user. Although our video clips are designed to reduce the cognitive load in remembering the passwords, they can not completely remove the load. Thus, our scheme is not scalable from the perspective of the users. We plan to conduct a multiple-password study in future to better assess this property.

3. *Nothing-to-Carry*: As users do not need to carry any additional physical object such as a piece of paper or a mechanical key, our scheme provides this benefit.

4. *Physically-Effortless*: A scheme is said to be Physically-Effortless if users do not require any additional physical effort beyond simply pressing a button to use the scheme. Since our scheme requires the users to type the password, it does not offer this benefit. However, our scheme does not require the users to type uppercase letters, or insert digits and special characters when constructing the password. We, therefore, consider our system to be Quasi-Physically-Effortless.

5. *Easy-to-Learn*: A scheme provides this benefit if first time users of the scheme can figure it out and learn it easily, as well can recall later how to use it. Our scheme involves watching a simple video clip and during our user study we observed

that users had no trouble in following it. Thus, we believe that our scheme is an Easy-to-Learn one.

6. *Efficient-to-Use*: For achieving this property, a scheme needs to have an acceptably short login time. Our results demonstrated that the median login times were 6 and 9 seconds, respectively, for the *link method*, and the *method of loci*. Furthermore, the maximum duration of the one-time registration activity is 180 seconds or 3 minutes. For these reasons, we claim that our scheme is an Efficient-to-Use one.

7. *Infrequent-Errors*: This property ensures that the false reject rate is low and genuine users are not frequently rejected. Since typos are always associated with password typing, our scheme is Quasi-Infrequent-Errors [11].

8. *Easy-Recovery-from-Loss*: In our scheme, if a user forgets her password, it can be easily reset. However, another video clip is required to be shown if a new password is to be assigned. This essentially grants a Quasi-Easy-Recovery-from-Loss status to our scheme.

Deployability Benefits

1. *Accessible*: In their paper, Bonneau et al. defined this benefit with respect to traditional textual password-based authentication scheme [11]. Users are not prevented from using the traditional password scheme due to disabilities or other physical (not cognitive) conditions, and any scheme that offers the same benefit is said to be Accessible. With regard to this notion, our system is not Accessible because visually impaired persons are not able to use our scheme. Such users can use regular websites with braille keyboards and interfaces.

2. *Negligible-Cost-per-User*: Since our scheme does not require any additional major hardware or graphics requirements, it offers the benefit of Negligible-Cost-per-User.

3. *Server-Compatible*: A scheme is Server-Compatible if it is compatible with text-based passwords at the verifier's end. The backend database for our scheme is essentially the same as the system-assigned textual password scheme. However, in case of our scheme, the registration server needs to provide users with the video, which is generated in part on the fly. We, therefore, consider our scheme to be Quasi-Server-Compatible.

4. *Browser-Compatible*: A scheme is said to be Quasi-Browser-Compatible if it requires any non-standard but very common browser plugins, such as Flash. In this regard, our scheme is Quasi-Browser-Compatible since it requires Unity plugin.

5. *Mature*: A scheme achieves this status once it has been deployed on a large scale in real world. As our proposed method is at its early stage and has just undergone the laboratory user study phase, it has yet to achieve this status.

6. *Non-Proprietary*: We assure that the proposed method will never be patented and always remain a Non-Proprietary scheme.

Security Benefits

1. *Resilient-to-Physical-Observation*: Since our scheme is not resilient against shoulder surfing attacks, it is not Resilient-to-Physical-Observation. However, Tari et al. showed that passwords using keyboard inputs provide higher resilience to shoulder surfing than passwords using mouse inputs, e.g., graphical passwords [114].

2. *Resilient-to-Targeted-Impersonation*: Since the password is assigned randomly by the system, any acquaintance or skilled investigator is unable to impersonate a user in our scheme by leveraging personal information such as birth date or high school name. Thus, our scheme is Resilient-to-Targeted Impersonation.

3. *Resilient-to-Throttled-Guessing*: Since the password is assigned randomly by the system in our scheme, it is resistant against throttled guessing, where an attacker's rate of guessing is constrained by the verifier.

4. *Resilient-to-Unthrottled-Guessing*: The entropy used in our study adheres to the prior study on system-assigned passwords [123]. However, it may not satisfy the benchmark suggested by Bonneau et al. to provide resilience against unthrottled-guessing [11].

5. *Resilient-to-Internal-Observation*: As our scheme is susceptible to keylogging malware, it is not Resilient-to-Internal-Observation.

6. *Resilient-to-Leak-from-Other-Verifiers*: This benefit ensures that even if a verifier leaks a password file to an attacker, the attacker would still be unable to impersonate the user to another verifier. This type of attack works because users tend to reuse the same password for multiple sites in the traditional user-chosen password scheme [51]. In case of our scheme, the password is assigned by the system. As a result, passwords can not be reused between systems that both use our scheme. However, users can memorize our assigned password and reuse this password in other systems that allow them to choose their own passwords. From this viewpoint, our scheme can be considered as Quasi-Resilient-to-Leak-from-Other-Verifiers.

7. *Resilient-to-Phishing*: Since phishing still remains an open problem in real-world, it would be unfair to claim that our scheme is resilient to phishing.

8. *Resilient-to-Theft*: As our scheme requires no physical object for authentication, it is Resilient-to-Theft.

9. *No-Trusted-Third-Party*: Our scheme provides this benefit since it does not depend on a trusted third party.

10. *Requiring-Explicit-Consent*: Schemes that initiate the authentication process without the explicit consent of the users fail to offer this benefit. Having to type the passwords makes our scheme Requiring-Explicit-Consent.

11. *Unlinkability*: This benefit can be obtained in our scheme if the sites add password salt independently.

In short, our scheme offers 17 of the 25 desired benefits, either completely, or in forms of quasi-benefit.

6.7 Memorizing Cryptographically-Strong Passwords by Leveraging the Method of Loci

Since the lab study results show promise for the *method of loci* in offering password-level security, it seems reasonable to examine its efficacy for providing crypto-level security [9]. To this end, we conducted a separate study.

We first discuss cryptographically-strong passwords in this section.

6.7.1 Cryptographically-Strong Passwords

In their work, Biddle et al. identify three specific ranges for theoretical password space and classify passwords into three categories accordingly: passwords that provide 20 bits of entropy (PIN-level security), passwords that provide 20 to 60 bits of entropy (password-level security), and passwords that provide more than 60 bits of entropy (crypto-level security) [9]. Since the passwords of this final category offer crypto-level security, they are also known as cryptographically-strong passwords.

Although the idea of cryptographically-strong passwords seems farfetched for regular Web authentication, this type of passwords are useful for several other important authentication applications, such as enterprise account login, master password for password managers, and password for protecting private keys in cryptography.

6.7.2 Spaced Repetition for Memorizing Cryptographically-Strong Passwords

In general, the security community tends to undermine the capability of human memory and the conventional wisdom is that users are not capable of remembering cryptographically-strong secrets [63]. In their work, Bonneau et al. challenge this notion and demonstrate that the *spaced repetition* technique is quite effective to imprint a 56-bit password into users' long-term memory [13].

In their study, Bonneau et al. assigned each participant a random 56-bit security code, which was represented as three chunks of four lowercase letters. For memorizing this code, a participant had to log into a website 90 times over up to 15 days. During the first login, the first chunk was displayed directly. For each subsequent login, a $\frac{1}{3}$ second delay was added before displaying the chunk, for the purpose of encouraging the participant to type the chunk from memory to save time. This delay was increased up to a maximum of 10 seconds. Once the participant was able to enter the first chunk before it was displayed, the same procedure was followed for the second and third chunks.

Three days after the last login, a follow-up study was arranged where the participant was asked to recall the code from the memory. Participants returned after a median of 3 days 18 hours for this study. Overall, 82% of participants were able to recall the code correctly from the memory.

We observe that although the study results are promising and more than four-fifth of the participants recalled the code successfully, the time spent for memorizing the secret is very long. The requirement of logging into a website 90 times over up to 15 days does not reflect the real-life time constraint, even for learning any important authentication secret. We believe that by leveraging the *method of loci*, the same goal can be achieved within a significantly shorter time period. In particular, our goal is to

leverage the *method of loci* to help users memorize a cryptographically-strong secret in just a single session.

We note that a password containing twelve lowercase letters yields an entropy value of 56.4 bits (Section 6.1.1), which is slightly lower than 60 bits. According to the range specified by Biddle et al. [9], it can not be considered as a cryptographically-strong secret. However, as we want to compare our scheme with that of Bonneau et al., we use the same level of entropy as used in their study. Moreover, in their work, Bonneau et al. consider a 56-bit code to be a cryptographically-strong secret [13].

6.8 Method of Loci for Memorizing Cryptographically-Strong Passwords

For our previous study with the *method of loci*, we used a virtual apartment model consisting of a mail box, a living room, a kitchen, a dining room, a bedroom, and a washroom. For this study, we extended this model and added six more loci which corresponded to six different rooms of a virtual office: a reception room, a file cabinet room, a copier room, a room of cubicles, a recreation room, and a conference room.

As cryptographically-strong passwords are used for high-value applications (Section 6.7.1), users have much incentive to spend more time in memorizing these passwords. We, therefore, relaxed the time constraint for password memorization in this study. The objects were displayed at the specified loci for a much longer period of time. The duration of the entire video clip for this study was 480 seconds or 8 minutes.

6.8.1 Participants, Apparatus, Environment and Procedure

For this experiment, we recruited 17 undergraduate and graduate participants from UTA. Participants came from diverse backgrounds, including majors from En-

gineering, Business, Psychology, and Biology. The age of the participants varied between 19 to 41 with a mean age of 24. In exchange of their time, the participants received a gift card for the Subway restaurant with a value of 10 dollars. The experiment consisted of two sessions and just like the previous experiment, the second session took place one week after the first one.

6.8.1.1 Session 1

In the first session, the participants were given an overview of the study after signing a consent form. Next they were assigned a random password with twelve lowercase letters and shown a video clip which was generated based on that password. After the clip ended, they were asked to write down sequentially the name of the twelve objects that were shown in the clip. If they forgot or missed the order of any object, the portion of the video clip displaying that object was shown again. Finally, they were shown the twelve objects in the specified loci for one last time.

After learning the password, participants logged into a site (designed by us using the images and layouts of a familiar commercial site) using the assigned password. We did not collect data for these practice trials. They were asked to not record (e.g., write down) the password. Before leaving, we reminded each participant to show up for the second part of the study after a week.

6.8.1.2 Session 2

After a week, when the participants returned for the second session, they were asked to log into the site using the assigned password. They were allowed to make a maximum of three attempts for a successful login. Participants who failed to recall correctly within three attempts were then provided a password hint. We showed them a video clip which navigated through the twelve loci, without displaying any

object. After that, they were given another chance to recall the password and log in successfully. Finally, all the participants were compensated and thanked for their time.

6.8.2 Results

As three of the participants did not return for the second session, we report the results for the remaining 14 participants.

6.8.2.1 Registration

The mean registration time was 737 seconds or 12 minutes 17 seconds.

6.8.2.2 Login

Our results show that out of 14 participants, 9 (64.3%) participants succeeded to log in by recalling the password correctly within three attempts. The mean login time was 36.73 seconds.

Of the remaining five participants, four were able to log in successfully after watching the password hint. The mean login time for these participants was 201 seconds, including the duration of the password hint clip (120 seconds).

Overall, out of the 14 participants who returned for the second session, only one participant failed to recall the password correctly even after watching the password hint clip. Thus, by incorporating the password hint, the login success rate for our scheme increases from 64% to 93%.

As mentioned before, Bonneau et al. conducted a similar study to test the effectiveness of spaced repetition in helping users to memorize 56-bit random passwords [13]. The recall success rate for their study was 82%, where the participants were required to log into a website 90 times over up to 15 days. We achieved a higher

recall success rate with just one training session. Although the sample size of our study is much smaller than that of Bonneau et al. (14 compared to 56), our results show promising initial evidence of the efficacy of the *method of loci* in assisting users to memorize cryptographically-strong secrets and suggest that it could be used as an alternative to the spaced repetition technique in this regard.

6.9 Conclusion

This is the first study to date that we are aware of which applies the *method of loci* to help users memorize a system-assigned authentication secret. Although both of our proposed methods have been used before in other domains, the current study is the first of its kind that leverages these methods to implement a training interface which does not require the users to do anything unaided, such as imagining some pictures of their own preferences or creating their own stories, to facilitate the memorization process. As a result, it is not susceptible to any kind of poor user action or selection.

Since our training method involves watching a video clip only, it is very simple to follow. We also try our best to ensure the desired security, usability, and deployability benefits of a good authentication mechanism. We offer sufficient password entropy, the duration of our training method is reasonable, and no overhead is associated with designing the login interface.

Our experimental results showed that both of our proposed methods outperformed the control condition with regard to login success rate. In fact, the *method of loci* had a login success rate of 86%, which is highest for any recall-based memorability study with system-assigned random passwords. Furthermore, the median login time was just 9 seconds and users expressed high satisfaction with the training interface for the *method of loci*.

We further extended the *method of loci* and leveraged it to help users memorize long passwords that offer almost crypto-level security. We conducted a separate experiment in this regard and found that with the help of a password hint, 93% of the participants were able to recall the password after a week. Thus, the *method of loci* provides a better alternative to spaced repetition technique for memorizing cryptographically-strong secrets, as it does not have the requirement of learning over a large number of sessions.

In future, we plan to observe the effect of memory interference for the *method of loci*. Memory interference is the impaired ability to remember an item when similar items are already stored in memory [5]. To observe the interference effect, we will conduct a multiple-password study with this method in future.

Finally, since our scheme does not require the users to type any uppercase letters, digits, or special characters, it offers a potential solution to the textual password entry problem on mobile devices [60]. Prior works have demonstrated the inconvenience of capitalizing letters and inserting digits or special characters when constructing a password by using a mobile device [48, 49]. Thus, we would like to implement and test our lowercase letter-only scheme on mobile platform in future.

REFERENCES

- [1] A. Adams and M. A. Sasse. Users are not the enemy. *Commun. ACM*, 42(12):40–46, 1999.
- [2] A. Adams, M. A. Sasse, and P. Lunt. Making passwords secure and usable. In *HCI*, 1997.
- [3] M. N. Al-Ameen and M. Wright. Multiple-password interference in the geopass user authentication scheme. In *USEC*, 2015.
- [4] M. N. Al-Ameen, M. Wright, and S. Scielzo. Towards making random passwords memorable: Leveraging users’ cognitive ability through multiple cues. In *CHI*, 2015.
- [5] M. Anderson and J. Neely. *Memory: Handbook of Perception and Cognition, Chapter 8: Interference and Inhibition in Memory Retrieval*. Academic Press, New York, 1996.
- [6] R. C. Atkinson and R. M. Shiffrin. Human memory: A proposed system and its control processes. In K. W. Spence and J. T. Spence, editors, *The Psychology of Learning and Motivation*. Academic Press, New York, 1968.
- [7] P. B. Baltes and U. Lindenberger. On the range of cognitive plasticity in old age as a function of experience: 15 years of intervention research. *Behavior Therapy*, 19:283–300, 1988.
- [8] P. Bao, J. Pierce, S. Whittaker, and S. Zhai. Smart phone use by non-mobile business users. In *MobileHCI*, 2011.
- [9] R. Biddle, S. Chiasson, and P. van Oorschot. Graphical passwords: Learning from the first twelve years. *ACM Computing Surveys*, 44(4), 2012.

- [10] M. Bishop and D. V. Klein. Improving system security via proactive password checking. *Computers & Security*, 14(3):233–249, 1995.
- [11] J. Bonneau, C. Herley, P. C. v. Oorschot, and F. Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *IEEE S & P*, 2012.
- [12] J. Bonneau and S. Preibusch. The password thicket: technical and market failures in human authentication on the web. In *The Ninth Workshop on the Economics of Information Security*, June 2010.
- [13] J. Bonneau and S. Schechter. Towards reliable storage of 56-bit secrets in human memory. In *USENIX*, 2014.
- [14] M. B. Brewer. Research design and issues of validity. In *H. T. Reis and C. M. Judd (Eds.), Handbook of research methods in social and personality psychology*, pages 3–16. New York: Cambridge University Press, 2000.
- [15] S. A. Brewster and M. Hughes. Pressure-based text entry for mobile devices. In *MobileHCI*, 2009.
- [16] J. D. Brown. *Testing in language programs*. Prentice Hall Regents, Upper Saddle River, NJ, 1996.
- [17] M. W. Browne and R. Cudeck. Alternative ways of assessing model fit. In *Testing structural equation models*. Sage Publications, Newbury Park, CA, 1993.
- [18] N. Burgess, E. A. Maguire, and J. O’Keefe. The human hippocampus and spatial and episodic memory. *Neuron*, 35(4):625–641, 2002.
- [19] M. D. C. Castelluccia and D. Perito. Adaptive password-strength meters from markov models. In *NDSS*, 2012.
- [20] H.-Y. Chiang and S. Chiasson. Improving user authentication on mobile devices: A touchscreen graphical password. In *MobileHCI*, 2013.

- [21] 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 *CCS*, 2009.
- [22] S. Chiasson, E. Stobert, R. Biddle, and P. van Oorschot. Persuasive cued click-points: Design, implementation, and evaluation of a knowledge- based authentication mechanism. *IEEE TDSC*, 9, 2012.
- [23] S. Chiasson, P. C. van Oorschot, and R. Biddle. Graphical password authentication using cued click points. In *ESORICS*, 2007.
- [24] J. P. Chin, V. A. Diehl, and K. L. Norman. Development of an instrument measuring user satisfaction of the human-computer interface. In *CHI*, 1988.
- [25] G. A. Churchill. A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, 16(1):64–73, February 1979.
- [26] M. Ciampa. A comparison of password feedback mechanisms and their impact on password entropy. *Information Management & Computer Security*, 21, 2013.
- [27] D. L. Clason and T. J. Dormody. Analyzing data measured by individual likert-type items. *Journal of Agricultural Education*, 35(4), 1994.
- [28] J. Cohen. *Statistical power analysis for the behavioral sciences (2nd ed.)*. Lawrence Erlbaum, Hillsdale, NJ, 1988.
- [29] L. J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, September 1951.
- [30] D. Davis, F. Monroe, and M. K. Reiter. On user choice in graphical password schemes. In *USENIX*, 2004.
- [31] X. de Carn de Carnavalet and M. Mannan. From very weak to very strong: Analyzing password-strength meters. In *NDSS*, 2014.

- [32] M. J. Dickel. Principles of encoding mnemonics. *Perceptual and motor skills*, 57:111–118, 1983.
- [33] A. C. Doyle. *The Complete Sherlock Holmes*. Barnes & Noble, New York City, 2009.
- [34] P. Dunphy and J. Yan. Do background images improve “Draw a Secret” graphical passwords? In *CCS*, 2007.
- [35] S. Fahl, M. Harbach, Y. Acar, and M. Smith. On the ecological validity of a password study. In *SOUPS*, 2013.
- [36] D. Florêncio and C. Herley. A large-scale study of web password habits. In *WWW*, 2007.
- [37] D. Florêncio and C. Herley. Where do security policies come from? In *SOUPS*, 2010.
- [38] A. Forget. *A world with many authentication schemes*. PhD thesis, Carleton University, 2012.
- [39] A. Forget, S. Chiasson, P. V. Oorschot, and R. Biddle. Improving text passwords through persuasion. In *SOUPS*, 2008.
- [40] P. Gasti and K. B. Rasmussen. On the security of password manager database formats. In *ESORICS*, 2012.
- [41] S. Gaw and E. W. Felten. Password management strategies for online accounts. In *Proceedings of the second symposium on Usable privacy and security*, pages 44–55, July 2006.
- [42] D. George and P. Mallery. *SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.)*. Allyn & Bacon, Boston, 2003.
- [43] E. E. Ghiselli. *Theory of psychological measurement*. McGraw-Hill, New York, 1964.

- [44] J. A. Gliem and R. R. Gliem. Calculating, interpreting, and reporting cronbach’s alpha reliability coefficient for likert-type scales. In *Midwest Research to Practice Conference in Adult, Continuing, and Community Education*, 2003.
- [45] J. Goldberg, J. Hagman, and V. Sazawal. Doodling our way to better authentication. In *CHI Extended Abstracts*, 2002.
- [46] J. S. Grant and L. L. Davis. Selection and use of content experts for instrument development. *Research in Nursing & Health*, 20(3):269–274, June 1997.
- [47] R. M. Guion. On Trinitarian doctrines of validity. *Professional Psychology*, 11(3):385–398, June 1980.
- [48] S. M. T. Haque, S. Scielzo, and M. Wright. Applying psychometrics to measure user comfort when constructing a strong password. In *SOUPS*, 2014.
- [49] S. M. T. Haque, M. Wright, and S. Scielzo. Passwords and interfaces: Towards creating stronger passwords by using mobile phone handsets. In *SPSM*, 2013.
- [50] S. M. T. Haque, M. Wright, and S. Scielzo. Passwords and interfaces: Towards creating stronger passwords by using mobile phone handsets. In *SPSM*, 2013.
- [51] S. M. 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.
- [52] J. Harry N. Boone and D. A. Boone. Analyzing likert data. *Journal of Extension*, 50(2), April 2012.
- [53] J. Hattie and R. W. Cooksey. Procedures for assessing the validities of tests using the “known-groups” method. *Applied Psychological Measurement*, 8(3):295–305, July 1984.
- [54] E. Hayashi and J. I. Hong. A diary study of password usage in daily life. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 2627–2630, May 2011.

- [55] C. Herley, P. C. Oorschot, and A. S. Patrick. Passwords: If we’re so smart, why are we still using them? In *FC*, 2009.
- [56] P. G. Inglesant and M. A. Sasse. The true cost of unusable password policies: Password use in the wild. In *CHI*, 2010.
- [57] B. Ives, K. R. Walsh, and H. Schneider. The domino effect of password reuse. *Commun. ACM*, 47(4):75–78, 2004.
- [58] P. Jaferian, K. Hawkey, A. Sotirakopoulos, M. Velez-Rojas, and K. Beznosov. Heuristics for evaluating IT security management tools. In *SOUPS*, 2011.
- [59] M. Jakobsson and R. Akapivat. Rethinking passwords to adapt to constrained keyboards. In *MoST*, 2012.
- [60] M. Jakobsson, E. Shi, P. Golle, and R. Chow. Implicit authentication for mobile devices. In *HotSec*, 2009.
- [61] H. F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, September 1958.
- [62] H. F. Kaiser. An index of factorial simplicity. *Psychometrika*, 39(1):31–36, March 1974.
- [63] C. Kaufman, R. Perlman, and M. Speciner. *Network security: Private communication in a public world*. Prentice Hall Press, New Jersey, 2002.
- [64] J. J. Kaye. Self-reported password sharing strategies. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 2619–2622, May 2011.
- [65] 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 *IEEE S&P*, 2012.

- [66] J. Kirakowski and B. Cierlik. Measuring the usability of web sites. In *HFES*, 1988.
- [67] J. Kirakowski and M. Corbett. Sumi: The software usability measurement inventory. *British Journal of Educational Technology*, 24(3):210–212, September 1993.
- [68] 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 *CHI*, 2011.
- [69] 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 *CHI*, 2011.
- [70] U. Konradt, H. Wandke, B. Balazs, and T. Christophersen. Usability in online shops: Scale construction, validation and the influence on the buyers’ intention and decision. *Behaviour & Information Technology*, 22(3):165–174, May-June 2003.
- [71] C. H. Lawshe. A quantitative approach to content validity. *Personnel Psychology*, 28(4):563–575, December 1975.
- [72] E. L. G. Legge, C. R. Madan, E. T. Ng, and J. B. Caplan. Building a memory palace in minutes: Equivalent memory performance using virtual versus conventional environments with the Method of Loci. *Acta Psychologica*, 141:380–390, 2012.
- [73] A. N. Leontev. *Activity, Consciousness, Personality*. Prentice Hall, Englewood Cliffs, NJ, 1978.
- [74] J. R. Lewis. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human Computer Interaction*, 7(1):57–78, January 1995.

- [75] Z. Li, Q. Sun, Y. Lian, and D. D. Giusto. An association-based graphical password design resistant to shoulder-surfing attack. In *ICME*, 2005.
- [76] H. Lorayne and J. Lucas. *The memory book*. Stein and Day, New York, 1974.
- [77] E. A. Maguire, E. R. Valentine, J. M. Wilding, and N. Kapur. Routes to remembering: The brains behind superior memory. *Nature Neuroscience*, 6(1):90–95, 2003.
- [78] J. P. McIver and E. G. Carmines. *Unidimensional scaling*. Sage, Beverly Hills, CA, 1981.
- [79] N. McNamara and J. Kirakowski. Defining usability: quality of use or quality of experience? In *IPCC*, 2005.
- [80] S. Moffat and N. Hurran. *Sherlock*. BBC One, 2014.
- [81] P. E. Morris and R. Stevens. Linking images and free recall. *Journal of Verbal Learning and Verbal Behavior*, 13(3):310–315, 1974.
- [82] J. Nicholson, L. Coventry, and P. Briggs. Age-related performance issues for PIN and face-based authentication systems. In *CHI*, 2013.
- [83] G. Notoatmodjo and C. Thornborson. Passwords and perceptions. In *Proceedings of the Seventh Australasian Conference on Information Security - Volume 98*, pages 71–78, January 2009.
- [84] J. C. Nunnally. *Psychometric theory (1st ed.)*. McGraw-Hill, New York, 1967.
- [85] J. C. Nunnally. *Psychometric theory (2nd ed.)*. McGraw-Hill, New York, 1978.
- [86] J. C. Nunnally and I. H. Bernstein. *Psychometric theory (3rd ed.)*. McGraw-Hill, New York, 1994.
- [87] M. T. Orne. Demand characteristics and the concept of quasi-controls. In *Artifact in Behavioral Research*, R. Rosenthal and R.L. Rosnow (eds.), Academic Press, New York. 1969.

- [88] A. Paivio. *Imagery and verbal processes*. Holt, Rinehart and Winston, New York, 1971.
- [89] A. Parasuraman, V. A. Zeithaml, and L. L. Berry. SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, 64(1):12–40, Spring 1988.
- [90] R. Parasuraman and M. Rizzo. *Neuroergonomics: The brain at work*. Oxford University Press, New York, 2008.
- [91] N. Patel, J. Clawson, and T. Starner. A model of two-thumb chording on a phone keypad. In *MobileHCI*, 2009.
- [92] D. Pavlas, F. Jentsch, E. Salas, S. M. Fiore, and V. Sims. The play experience scale: Development and validation of a measure of play. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 54(2):214–225, April 2012.
- [93] S. Preibusch and J. Bonneau. The password game: negative externalities from weak password practices. In *Proceedings of the First international conference on Decision and game theory for security*, pages 192–207, November 2010.
- [94] A. Raz, M. G. Packard, G. M. Alexander, J. T. Buhle, H. Zhu, S. Yu, and B. S. Peterson. A slice of pi : An exploratory neuroimaging study of digit encoding and retrieval in a superior memorist. *Neurocase*, 15(5):361–372, 2009.
- [95] J. T. E. Richardson. The efficacy of imagery mnemonics in memory remediation. *Neuropsychologia*, 33:1345–1357, 1995.
- [96] J. Robertson. Stats: We’re doing it wrong. <http://cacm.acm.org/blogs/blog-cacm/107125-stats-were-doing-it-wrong/fulltext>, April 2011.
- [97] H. L. Roediger. The effectiveness of four mnemonics in ordering recall. *Journal of Experimental Psychology: Human Learning and Memory*, 6(5):558–567, 1980.

- [98] J. R. Rossiter. The C-OAR-SE method and why it must replace psychometrics. *European Journal of Marketing*, 45(11):1561–1588, November 2011.
- [99] Y. S. Ryu and T. L. Smith-Jackson. Reliability and validity of the Mobile Phone Usability Questionnaire (MPUQ). *Journal of Usability Studies*, 2(1):39–53, November 2006.
- [100] M. A. Sasse, S. Brostoff, and D. Weirich. Transforming the ‘weakest link’- A human/computer interaction approach to usable and effective security. *BT Technology Journal*, 19(3):122–131, 2001.
- [101] F. Schaub, M. Walch, B. Konings, and M. Weber. Exploring the design space of graphical passwords on smartphones. In *SOUPS*, 2013.
- [102] F. Schaub, M. Walch, B. Konings, and M. Weber. Exploring the design space of graphical passwords on smartphones. In *SOUPS*, 2013.
- [103] B. Schneier. *Secrets and Lies : Digital Security in a Networked World*. Wiley Computer Publishing, New York, 2004.
- [104] C. A. Schriesheim, K. J. Powers, T. A. Scandura, C. C. Gardiner, and M. J. Lankau. Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments. *Journal of Management*, 19(2):385–417, April 1993.
- [105] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.
- [106] 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 *SOUPS*, 2012.
- [107] R. Shay, S. Komanduri, P. G. Kelley, P. G. Leon, M. L. Mazurek, L. Bauer, N. Christin, and L. F. Cranor. Encountering stronger password requirements:

- user attitudes and behaviors. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, July 2010.
- [108] R. Shay, S. Komanduri, P. G. Kelley, P. G. Leon, M. L. Mazurek, L. Bauer, N. Christin, and L. F. Cranor. Encountering stronger password requirements: User attitudes and behaviors. In *SOUPS*, 2010.
 - [109] S. W. Smith. Humans in the loop: Human-computer interaction and security. *IEEE Security & Privacy*, 1, 2003.
 - [110] C. Spearman. *The abilities of man*. Macmillan, New York, 1927.
 - [111] P. Spector. *Summated rating scale construction*. Sage, Thousand Oaks, CA, 1992.
 - [112] B. Stock and M. Johns. Protecting users against xss-based password manager abuse. In *ASIA CCS*, 2014.
 - [113] H. Tang, D. J. Beebe, and A. F. Cramer. A multilevel input system with force-sensitive elements. *International Journal of Human-Computer Studies*, 54(4):495–507, April 2001.
 - [114] F. Tari, A. Ozok, and S. Holden. A comparison of perceived and real shoulder-surfing risks between alphanumeric and graphical passwords. In *SOUPS*, 2006.
 - [115] J. Thorpe, B. MacRae, and A. Salehi-Abari. Usability and security evaluation of geopass: A geographic location-password scheme. In *SOUPS*, 2013.
 - [116] E. Tulving. Cue-dependent forgetting: When we forget something we once knew, it does not necessarily mean that the memory trace has been lost; it may only be inaccessible. *American Scientist*, 62(1):74–82, 1974.
 - [117] 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 *USENIX*, 2012.

- [118] M. Weir, S. Aggarwal, M. Collins, and H. Stern. Testing metrics for password creation policies by attacking large sets of revealed passwords. In *CCS*, 2010.
- [119] A. Whitten and J. D. Tygar. Why Johnny can't encrypt: A usability evaluation of PGP 5.0. In *USENIX*, 1999.
- [120] S. Wiedenbeck, J. Waters, J.-C. Birget, A. Brodskiy, and N. Memon. Authentication using graphical passwords: Effects of tolerance and image choice. In *SOUPS*, 2005.
- [121] S. Wiedenbeck, J. Waters, J.-C. Birget, A. Brodskiy, and N. Memon. Pass-Points: Design and longitudinal evaluation of a graphical password system. *International Journal of Human-Computer Studies*, 63(1-2), 2005.
- [122] D. Wigdor and R. Balakrishnan. A comparison of consecutive and concurrent input text entry techniques for mobile phones. In *CHI*, 2004.
- [123] N. Wright, A. S. Patrick, and R. Biddle. Do you see your password?: Applying recognition to textual passwords. In *SOUPS*, 2012.
- [124] J. Yan, A. Blackwell, R. Anderson, and A. Grant. Password memorability and security: Empirical results. *IEEE Security & Privacy*, 2(5):25–31, 2004.
- [125] F. A. Yates. *The art of memory*. University of Chicago Press, Chicago, 1966.
- [126] J. A. Yesavage. Imagery pretraining and memory training in the elderly. *Gerontology*, 29:271–275, 1983.
- [127] D. A. Z. Li, W. He and D. Song. The emperor's new password manager: Security analysis of web-based password managers. In *USENIX Security*, 2014.
- [128] Y. Zhang, F. Monrose, and M. K. Reiter. The security of modern password expiration: An algorithmic framework and empirical analysis. In *CCS*, 2010.
- [129] M. Zviran and W. J. Haga. A comparison of password techniques for multilevel authentication mechanisms. *The Computer Journal*, 36(3):227–237, 1993.

- [130] M. Zviran and W. J. Haga. Password security: An empirical study. *J. Manage. Inf. Syst.*, 15(4):161–184, 1999.

BIOGRAPHICAL STATEMENT

S M Taiabul Haque was born in Dhaka, Bangladesh, in 1985. He graduated with a bachelor's degree in Computer Science and Engineering from Bangladesh University of Engineering and Technology (BUET). He received his PhD degree from the Department of Computer Science and Engineering at the University of Texas at Arlington (UTA). Before moving to USA, he had worked as a full-time Lecturer in the Department of Computer Science and Engineering at State University of Bangladesh (SUB). His research interests lie in the areas of Information Security and Human-Computer Interaction. He is specifically interested in exploring the underlying human factors in textual password-based authentication system. He has contributed five peer-reviewed publications in this field. He has been invited by Cornell University Information Science Department to give a talk at their Distinguished Young Scholars Speaker Series, as well been cited in a federal government publication of National Institute of Standards and Technology. In Fall 2015, he will be joining the Department of Mathematics and Computer Science at the University of Central Missouri (UCM) as an Assistant Professor.