Text Entry Method Affects Password Security

Yulong Yang[†], Janne Lindqvist[†], Antti Oulasvirta[‡]
[†]Rutgers University, [‡]Aalto University

Abstract

Background. Text-based passwords continue to be the primary form of authentication to computer systems. Today, they are increasingly created and used with mobile text entry methods, such as touchscreen qwerty keyboards, in addition to traditional physical keyboards.
Aim. This paper aims to answer a foundational question for usable security: whether text entry methods affect password generation and password security.
Method. This paper presents results from a between-group study with 63 participants, in which each group generated passwords for multiple virtual accounts using different text entry methods. Participants were also asked to recall their passwords afterwards.
Results. One-way ANOVA across groups was performed on metrics including password length, amount of different characters, and estimated password security. The results showed significant effect of text entry methods on the amount of lowercase letters per password across groups ($F(2,60) = 3.186$, $p = .048$, $\eta_p^2 = .066$), and non-significant effect on the password length, amount of uppercase letters, digits or symbols. No significant result was found for the estimated password security. The result of practical cracking attacks was also similar across groups.
Conclusions. Text entry methods have effect on password security. However, the effect is subtler than expected.

1 Introduction and Background

Text-based passwords remain as the most prevalent method of authentication [1]. In addition to traditional computers such as desktops and laptops, people increasingly generate and use passwords with a wide variety of mobile terminals, such as tablets and smartphones. These mobile terminals have very different text entry methods compared to traditional computers. The appearance of such text entry methods drastically diversified how we use password authentication.

Mobile terminals are also replacing traditional computers in daily tasks. For example, Pew Research estimates that 21% of all US adult cell phone owners use primarily their phone to access the web [2].

A *text entry method* [3] consists of those physical (e.g. form factor, display, etc.) and software (e.g. virtual keyboard layout) aspects of an input device that are relevant when entering text. The design of a text entry method determines how quickly and effortlessly a given character can be typed. Even small changes in how characters are displayed and organized can affect typing performance [4]. As a result, experienced typists on physical keyboards reach more than 60 words per minute (wpm) [5], whereas tablets and smartphones are in the range of 20 to 30 wpm [6,7]. Further, one should see corresponding differences in the distribution of characters in different methods. For instance, digits are not directly reachable without changing the layout in the common touchscreen qwerty keyboard on smartphones – does this affect the generated passwords?

2 Aim

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

We first examined whether structure of generated passwords. Metrics of password structure included password length, the amount of lowercase letters, uppercase let-

ters, symbols and digits per password. We also looked at types of passwords. We defined the type of a password by types of characters it contained.

Second, we explored the question whether text entry methods affected password security. We estimated security of passwords from two aspects: quantitative estimation and practical cracking attacks. Quantitative estimation included Shannon entropy [8, 9], NIST entropy [10] and a recently introduced Markov-model-based metric (adaptive password-strength meter [11]). Then we looked at was how resistant passwords were against cracking attacks. We issued both dictionary attacks and rule-based guessing attacks.

Finally, we studied if participants perceived the task with different text entry methods differently using NASA Task Load Index assessment (TLX) [12].

3 Related Work

In this section, we first focus on what is known about generating passwords with mobile text entry methods, and then password generation in general.

Researchers have studied usability of mobile platforms for passwords. Greene et al. [13] studied the difference between typing passwords using tablet and smartphone in a between-group experiment. The time used for participants to type and recall passwords were significantly different provided with different entry methods. They were also different given different passwords. Schaub et al. [14] found similar significant difference among different smartphones. In addition, they found that attackers had significantly different success rates in shoulder surfing passwords on different smartphones. Both of mentioned studies did not ask participants to create passwords, and provided participants passwords instead, thusly having little information on password generation and consequently how password security would be affected if created using different entry methods.

Few studies have specifically looked at helping people to create passwords on mobile text entry methods. Haque et al. [15] have studied how to create better passwords on mobile devices. They found entropy of passwords were significantly different across mobile keyboards. However, only an approximation of Shannon entropy was examined. An analysis with other security metrics and also password structures could help us gain more insight on the effect of text entry methods. Jakobsson et al. [16] proposed fastwords, which relied on standard error-correcting features for users to create passphrases.

Florencio et al. [17] have studied web password habits in a large scale. They found that most people managed multiple passwords, and their passwords were generally of poor quality, and were re-used and forgotten frequently. Grawemeyer [18] conducted a diary study and found people had different strategies for different accounts. Such studies indicated that multi-account scenario would be reasonable in password experiments.

Recently, Bonneau et al. [19] studied how people chose 4-digit PINs for banking cards. Common strategies included birth dates and visual patterns. The reported presence of visual strategies supported our hypothesis that password generated with different text entry methods, too, may differ.

Researchers have also studied how password generation policies affected password security. Weir et al. [20] claimed that passwords created under common requirements, such as minimum length and different character set requirements, were still vulnerable to cracking attacks. Shay et al. [21] found that some policies that required longer passwords provided better usability and security compared with traditional policies. Ur et al. [22] found that stringently rated password meters led users to make significantly longer passwords that included more types of characters, and passwords were also more resistant against cracking algorithms. However, Egelman et al. [23] showed that password meters helped little if people considered the accounts unimportant.

Therefore, specific policies and requirements do affect password generation and security. To exclude such effect from our experiment, one might need to avoid explicit password generation requirements.

Finally, Fahl et al. [24] compared real passwords to those generated in an experiment, finding that about 30% of subjects do not behave as they do in real life. However, the authors concluded that laboratory studies generally created useful data.

4 Method

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

4.1 Experiment Design

The experiment followed a between-group design with text entry method type (3 levels) as independent variable.

We divided our participants into three groups based on text entry method variable. The participants were randomly assigned into one of these three groups, and were unaware of the assignments or that other groups existed. A detailed explanation for the differences between groups was given in the next subsection.

The main reason for us to choose between-group was to isolate its effect from any other undesired effects such as any possible confounding factors that would correlate

with both the variable and the result. We noticed that previous work that involved password generation process also had similar experiment design [13, 14, 25].

An alternative design would be within-subject, in which one participant would perform the same task using three different text entry methods in a sequence. In such design, the use of different text entry methods would generate undesired interference to each other for each participant. In particular, learning and using one text entry method would interfere with the learning and using of other text entry methods, thus decreasing or even eliminating the potential effect of both methods. Such interference is common in paired tasks [26, 27].

Florencio et al. revealed that people manage multiple passwords in reality [17]. To increase ecological validity, we asked participants to manage three different virtual accounts. However, since the difference within each participant was not in our research objective, we did not analyze the difference among three passwords created by each participant. Instead, mean value of three accounts was taken to represent each participant in our models.

4.2 Apparatus

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

Laptop group (control group)

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

Tablet group

We provided Samsung Nexus 10 tablet (Android 4.2.2, 10.1" touchscreen) as the device used in tablet group. The touchscreen keyboard on the tablet had a common qwerty layout, as shown in Figure 1. Given that the tablet can be held in the hands in two ways, we asked the participants to keep it in the "landscape" mode.

Smartphone group

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

The difference between our smartphone keyboard and tablet keyboard was the number of key presses needed to reach certain keys (see Figure 1). To reach uppercase letters, one needed to press two additional keys from the first layout in smartphone group, while only one key press in tablet group. Also, to reach special symbols layout, three additional key presses were needed in smartphone group while only two in tablet group. In short, reaching certain keys and switching layouts demanded more effort in smartphone keyboard than tablet keyboard. The primary reason we chose our text entry methods so was to differentiate each group in their difficulty of reaching keys during text entry. All three apparatus still provided common usability for the particular platform.

Software

The application was implemented in both Python (for laptop group) and Java for Android (for smartphone and tablet group). It had mainly two features: password creation and password recall. In the password creation interface, participants were asked to create usernames and passwords for three virtual accounts in the same order. Each virtual account had a different logo, color and short description. In the recall interface, it asked participants to recall what they created earlier for each account, in a different order. "Give up" button would show up after four failed attempts for each account.

4.3 Procedure

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

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

Our study consisted two sessions. In session one, we asked participants to create a username and a secure password for three different accounts given a certain text entry method. The detailed procedure was as follows:

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

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

3. **Subjective Workload Assessment** The participants were asked to fill out the NASA TLX form [12].

4. **Distraction**. The participants were asked to do a mental rotation task [28] and count down from 20 to 0 in mind.

5. **Password Recall**. Participants were asked to recall usernames and passwords they created in the Password Creation step above. The order of the accounts were changed with Latin square. For each account, participants were allowed to try as many time as

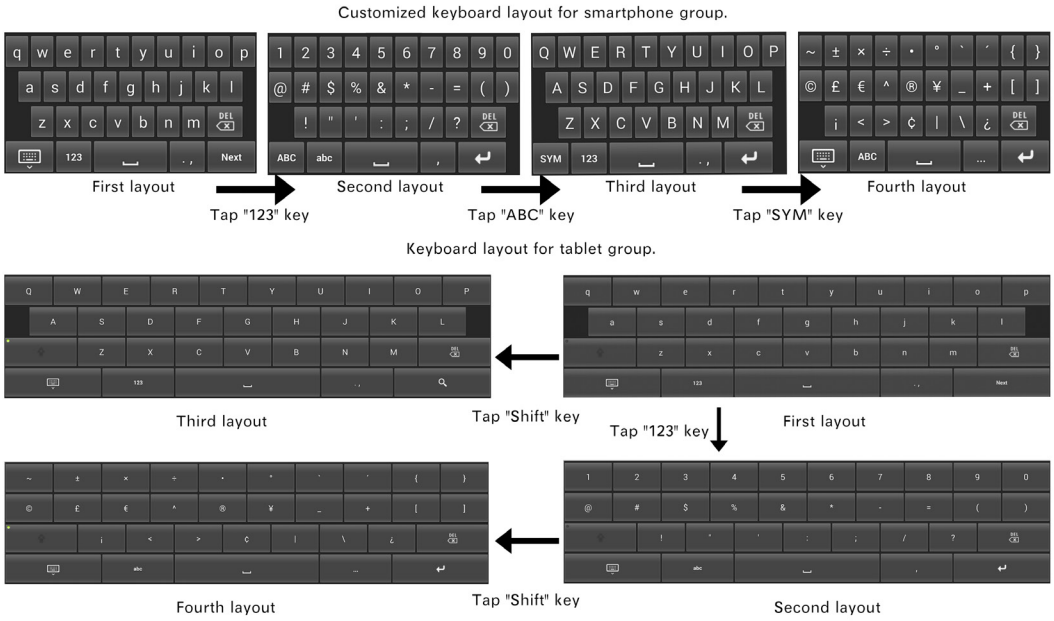

Figure 1: The keyboard layout for the devices in tablet group and smartphone group. Note two groups shared the same key positions within each layout, but the structures of the four layouts were different for them: tablet group followed the more common structure, while smartphone group had a hierarchical structure. To reach the next layout of smartphone keyboard, one had to first reach the previous one. Therefore, smartphone keyboard had a higher difficulty reaching non-lowercase keys than tablet keyboard.

they wanted, and give up if necessary (showed up after four failed attempts).

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

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

4.4 Participants

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

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

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

The number of participants for laptop group, tablet group and smartphone group were 21, 27 and 15.

Non-equal group sizes are expected after random assignment [29]. Tests applied in following sections were applied to the entire sample distribution. To ensure the validity of results, we randomly sampled our two larger groups so that the size was even across groups, and then performed same tests again. The result on sampled data were the same, indicating our tests were robust against the unbalanced group size.

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

4.5 Password security estimation

We describe our password security estimation below.

As a measure of "uncertainty", Shannon entropy had been used in evaluating security of passwords in cryptographic contexts [10]. We used random entropy in our analysis, which was defined as in equation $H = L \times \log_2 N$, in which L was the length of the password, and N was the possible set of characters.

The NIST entropy was a scheme to evaluate human-selected passwords introduced in NIST Electronic Authentication Guideline [10]. The scheme took into account the fact that passwords were chosen by human beings, who tend to choose passwords that were easily guessed, and even from a set of a few thousand commonly chosen passwords. We implemented the scheme by assigning different entropy to characters at different positions, each password creation rule contributing a specific amount of entropy and that the entropy of the policy was the sum of the entropy contributed by each rule. In addition, we performed a simple dictionary word check ("dic-0294") to give the password extra entropy.

The adaptive password-strength meter (APSM) based on Markov models estimated the strength of a password by estimating the probability of n-grams that composed the password [11]. N-gram is a contiguous sequence of n characters from a given string. Probabilities of n-grams are computed based on a large password dataset, therefore, it introduces certain dependency on the training password dataset. In our implementation, we used the "Rockyou" password dataset to compute the database of probabilities for every n-gram. The dataset contained over 32 million real passwords. We chose 4-gram as the element in our implementation as the original paper did.

There were some other metrics we did not include in our analysis. Bonneau has proposed several statistical metrics for password security [30]. However, Bonneau's metrics were mainly applicable to a large-scale password dataset, while we had a much smaller one.

4.6 Password Cracking attacks

We performed several actual cracking attacks against our passwords. We used two popular password cracking tools, John the Ripper[1] and hashcat[2].

Dictionaries

We used various dictionaries that were common in the literature. "dic-0294" was a English dictionary from outpost9 [3]. "All" was a free public dictionary from openwall website[4]. "Mangled" was a paid dictionary from openwall. It was a hand-tuned wordlist containing four million password candidates generated using various mangle rules. "Rockyou" included about 32 million passwords leaked from the website RockYou. "Facebook" was a list of names of searchable user from the website Facebook [31]. "Myspace" contained passwords from a phishing attack against MySpace website. "Inflection"[5] was a list of words along with their different grammatical forms such as plurals and past tense.

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

Dictionary attack

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

Long session offline attack

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

The first attack involved generating guesses based on a modified "Single mode" rules, which was originally from John the Ripper, using the "dic-0294" dictionary as input. The "Single mode" rules contained a set of rules to modify words including login names and directories to generate guesses [32]. The modified version, made by Weir [33], was optimized for English dictionary. We followed the same setup of Weir et al. [20].

The second attack applied the probability password crack tool developed by Weir et al. [20, 34]. It generated password guesses in the order determined by various rules derived from training sets. We used a similar model from experiment P4 conducted by Kelley et al. [35].

5 Results

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

[1] http://www.openwall.com/john/

[2] http://hashcat.net/hashcat/

[3] http://www.outpost9.com/files

[4] http://www.openwall.com/wordlists/

[5] http://wordlist.sourceforge.net

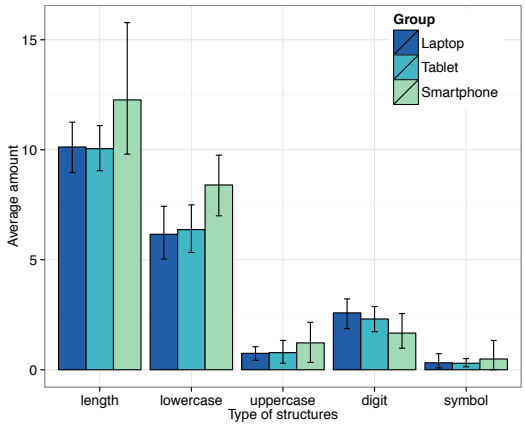

Figure 2: The average password length, amount of lowercase letters, uppercase letters, digits and symbols appeared in single password across groups. Error bars stand for 95% confidence intervals based on a bootstrap (that is, not assuming normality).

5.1 Structures

Figure 2 showed the password length and the amount of characters per password classified by types across groups. It demonstrated a notable difference in password length and amount of lowercase letters between smartphone group and other two groups.

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

Next, we examined the categories of passwords each group generated. We defined the category of a password by types of characters it contained. The category of a password revealed the complexity in its structures: passwords containing multiple types of characters had a more complex structure than ones with only one type. Table 1 summarized our definition of categories.

Figure 3 showed the distribution of passwords within the defined categories across groups. For smartphone group, passwords that contained only lowercase letters (*loweralpha*) was most common (31.1%). For other two groups, passwords containing only lowercase letters and digits (*loweralpha-num*) were the most common: 30.2% in laptop group and 38.2% in tablet group, respectively. In addition, there was no passwords containing lowercase letters, special symbols and digits (*loweralpha-special-num*) in smartphone group at all, while both other groups generated passwords in that category.

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

Table 1: Definition of each category of passwords. All types with low occurrence in our passwords were aggregated into "others" category.

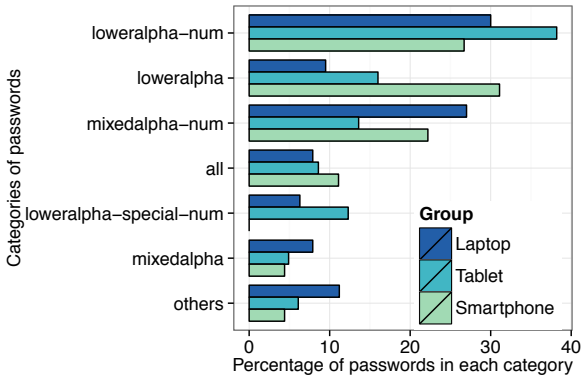

Figure 3: A comparison of distribution of passwords in different categories for each group.

5.2 Quantitative password security

We estimated the security of our passwords with two common entropy-based password security metrics, random entropy and NIST entropy, and a more recent Markov model based metric (APSM). Such metrics provided quantitative measurement of password security. Three metrics were explained in details in section 4.3. The mean scores and corresponding confidence intervals of the result were shown in Figure 4. According to the graph, scores of passwords of smartphone group were consistently higher than that of other two groups. However, most of means stayed within the confidence interval of the value of other groups, indicating the differences among groups were limited.

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

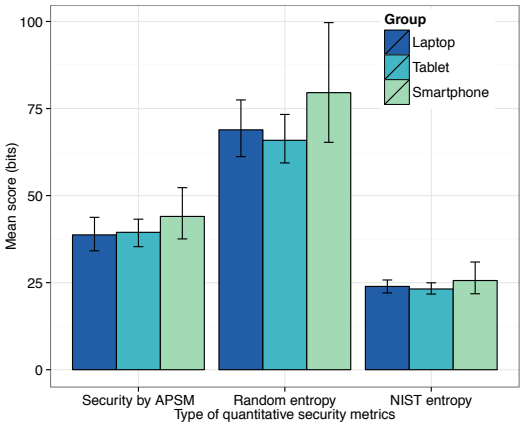

Figure 4: The mean score of three password security metrics across groups: score from Adaptive Password-Strength Meter (APSM), random entropy and NIST entropy. Error bars stand for 95% confidence intervals based on a bootstrap (that is, not assuming normality).

5.3 Cracking attacks

We performed dictionary attacks and long-session offline attacks on our collected passwords. Both attacks have been described in details in section 4.4. Table 2 showed the result of plain dictionary attacks. The performance of "Words" and "Facebook" attacks were limited across all groups, except "Facebook" attack on passwords in smartphone group. The "Password" attack worked much better compared to the first two attacks against laptop and tablet group, while it had very limited improvement over previous attacks against smartphone group.

The results of two long session offline attacks were shown in Figure 5 and Figure 6, respectively. According to the figures, although the lowerbounds of resistance (the number of guesses of the first cracked password) were different, the percentages of cracked passwords across groups were similar to each other.

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

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

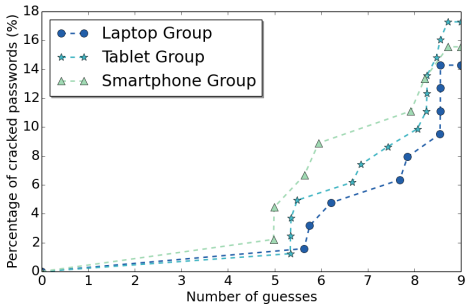

Figure 5: The percentage of passwords cracked by our first offline attack. The x-axis was in log scale. The final percentage of cracked passwords for laptop group, tablet group and smartphone group were 14.2%, 17.3% and 15.6%, respectively.

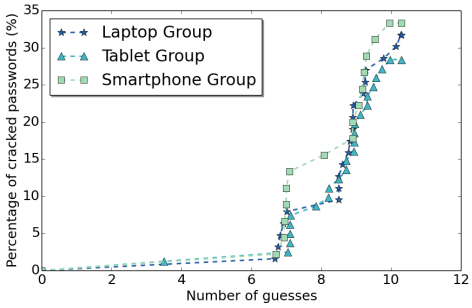

Figure 6: The percentage of passwords cracked by Weir's algorithm vs. the number of guess, per group. The x-axis was in log scale. The final percentage of cracked passwords for laptop group, tablet group and smartphone group are 31.7%, 28.4% and 30%, respectively.

5.4 Task load

We used TLX forms to evaluate the subjective task load of our study. These questions revealed participants' subjective assessment towards tasks in the study. Figure 8 showed the mean scores for each question of TLX form for both sessions.

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

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

Table 2: Results of both plain dictionary attacks and long-session offline attacks. “Include” listed all dictionaries we used in each attack. The size was the number of unique entries each combined dictionary had for dictionary attacks, and the number of guesses generated per password for long-session offline attacks.

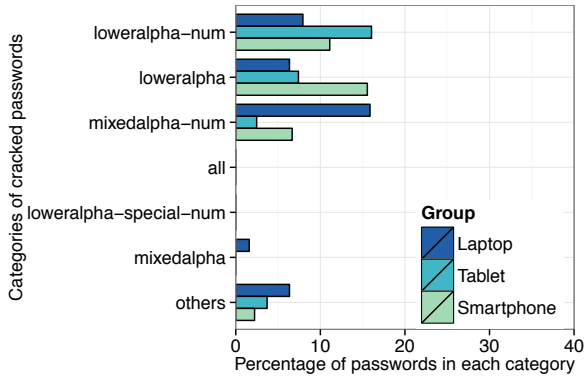

Figure 7: A comparison of percentages of cracked passwords in different categories across groups. The percentage value showed the percentage of cracked password in total amount of passwords in each group. We kept the categories and percentage scale as the same as in Figure 3 for better comparison. Cracked passwords here were the combination of cracked passwords in all our attacks.

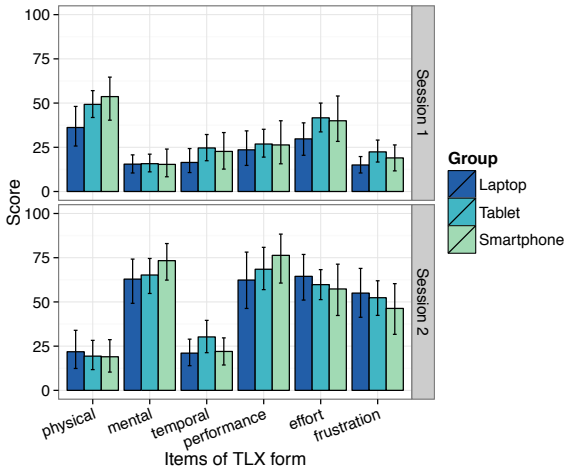

Figure 8: Mean score for each item in TLX form in session one and two. Error bars stand for 95% confidence intervals based on a bootstrap (that is, not assuming normality).

6 Discussion and Conclusions

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

One possible reason of such result could be that while passwords consisted more lowercase letters were considered weaker, smartphone group actually generated longest passwords in average (around 12.5, compared to 10 in other groups, see Figure 2). Extra length made passwords more secure. For example, a 15-character-long lowercase-only password from smartphone group scored 101, 28.5 and 48 in random entropy, NIST entropy and APSM, respectively. All of them are well above overall average.

In our study, smartphone keyboard demanded most effort in switching layouts. As a result, participants switched layouts less often in smartphone group, leading to more lowercase letters in passwords. However, participants still paid sufficient effort on passwords, resulting in long passwords. According to Shay et al. [21], long passwords were generally more secure. Meanwhile, smartphone group participants did not report a higher load in TLX forms (Figure 8).

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

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

Also, the study is conducted as a lab study, in which participants created passwords under the watch of exper-

imenters. It is possible that under such condition, participants spent extra effort to create passwords that are stronger than usual. For example, the average password length of each group in our study is at least 10, while that of RockYou passwords is below 8.

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

On the other hand, the analysis of password structure and cracking attacks still showed the effect existed. As mentioned before, the variable had significant effect on number of lowercase letters in passwords (Figure 2). This finding was consistent with our experiment design, as the difficulty of reaching non-lowercase keys in the smartphone group was increased. In addition, we found that passwords cracked in our attacks distributed quite differently in categories across groups (Figure 7). Particularly, nearly 50% of cracked passwords in every group belonged to a different single category compared with each other. It showed different resistance against cracking attacks across groups.

Limitations. Our sample size was relatively small, a large-scale study would be desired in the future. In addition, our study limited participants to create and recall passwords in a lab environment, which is not close to the real scenario when passwords are used. While recent study by Fahl et al. [24] showed that laboratory studies generally create useful data, a field study could be a follow-up on this topic. Also, while in our study we used common text entry methods, one could include more manipulations to see how would the effect be changed due to specific manipulations.

Conclusions. We presented the analysis of passwords created with different text entry methods. We designed and executed an experiment that aimed at exploring the possible effect of text entry methods on password security. Our results showed that the effect was not as significant as we hypothesized. The structure of passwords had been affected by such variable. However, it did not have significant effect on password security, according to our quantitative security estimation and cracking attacks. More work is needed to pinpoint the magnitude of the effect and exact design factors in text entry methods that affecting people's password generation process.

Acknowledgments

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

References

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

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

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

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

[5] D. R. Gentner, "The acquisition of typewriting skill," *Acta Psychologica*, 1983.

[6] M. Goel, L. Findlater, and J. Wobbrock, "Walktype: using accelerometer data to accomodate situational impairments in mobile touch screen text entry," in *Proc. of CHI'12*.

[7] A. Oulasvirta, A. Reichel, W. Li, Y. Zhang, M. Bachynskyi, K. Vertanen, and P. O. Kristensson, "Improving two-thumb text entry on touchscreen devices," in *Proc. of CHI'13*.

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

[9] J. Massey, "Guessing and entropy," in *Information Theory, 1994. Proceedings., 1994 IEEE International Symposium on*, 1994.

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

[11] C. Castelluccia, M. Düurmuth, and D. Perito, "Adaptive password-strength meters from markov models," in *Proc. of NDSS'12*, 2012.

[12] S. G. Hart and L. E. Staveland, *Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research*, 1988.

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

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

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

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

[17] D. Florencio and C. Herley, "A large-scale study of web password habits," in *Proc. of WWW'07*, 2007.

[18] B. Grawemeyer and H. Johnson, "Using and managing multiple passwords: A week to a view," *Interact. Comput.*, vol. 23, no. 3, 2011.

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

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

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

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

[23] S. Egelman, A. Sotirakopoulos, I. Muslukhov, K. Beznosov, and C. Herley, "Does my password go up to eleven?: the impact of password meters on password selection," in *Proc. of CHI'13*, 2013.

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

[25] S. Chiasson, A. Forget, E. Stobert, P. C. van Oorschot, and R. Biddle, "Multiple password interference in text passwords and click-based graphical passwords," in *Proc. of CCS'09*, 2009.

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

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

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

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

[30] J. Bonneau, "The science of guessing: Analyzing an anonymized corpus of 70 million passwords," in *Proc. of SP'12*, 2012.

[31] "Return of the facebook snatchers," 2010. [Online]. Available: https://blog.skullsecurity.org/2010/return-of-the-facebook-snatchers

[32] "John the ripper single crack mode." [Online]. Available: http://www.openwall.com/john/doc/MODES.shtml

[33] "Optimizing john the ripper's "single" mode for dictionary attacks," 2010. [Online]. Available: http://reusablesec.blogspot.com/2010/04/optimizing-john-rippers-single-mode-for.html

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

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