I Can't Type That! P@$$w0rd Entry on Mobile Devices*

Kristen K. Greene[1], Melissa A. Gallagher[2], Brian C. Stanton[1], and Paul Y. Lee[1]

[1] National Institute of Standards and Technology
100 Bureau Dr, Gaithersburg, MD, USA
{kristen.greene,brian.stanton,paul.lee}@nist.gov
[2] Rice University
6100 Main St, Houston, TX, USA
mg17@rice.edu

Abstract. Given the numerous constraints of onscreen keyboards, such as smaller keys and lack of tactile feedback, remembering and typing long, complex passwords — an already burdensome task on desktop computing systems —becomes nearly unbearable on small mobile touchscreens. Complex passwords require numerous screen depth changes and are problematic both motorically and cognitively. Here we present baseline data on device- and age-dependent differences in human performance with complex passwords, providing a valuable starting dataset to warn that simply porting password requirements from one platform to another (i.e., desktop to mobile) without considering device constraints may be unwise.

Keywords: Passwords, authentication, security, memory, mobile text entry, typing, touchscreens, smartphones, tablets.

1 Introduction

Despite widespread recognition that passwords are a fundamentally broken method of user authentication [1], they will almost certainly remain deeply embedded in today's digital society for quite some time. Unfortunately, the very features of a password that are intended to make it more secure (e.g., increasing length, use of mixed case, numbers, and special characters [2]) generally make it less usable. Remembering and typing long, complex passwords is already a burdensome task on desktop computing systems with full QWERTY keyboards; entering the equivalent text on mobile touchscreen devices will no doubt prove significantly more challenging for users. While this premise seems inarguable—especially given the numerous constraints of onscreen keyboards, such as smaller keys and lack of tactile feedback—it must nonetheless be supported by quantitative human data. Here we present baseline mobile data on device- and age-dependent differences in human performance with complex passwords, complementing the desktop study [3] upon which this work is based.

* The rights of this work are transferred to the extent transferable according to title 17 U.S.C. 105.

T. Tryfonas and I. Askoxylakis (Eds.): HAS 2014, LNCS 8533, pp. 160–171, 2014.

2 Text Entry

Text entry on mobile devices is a common subroutine in many tasks. Past work has examined the effect of different technologies [4], age [5], motion [6], and a number of different devices [7] [8] when participants are typing words or phrases. While other research (e.g., [9], [10], [11]) has examined non-word strings of random letters, such research did not include the variety of numbers and special characters recommended for passwords, neither for desktop nor for mobile devices. As both the number of accounts users interact with on their mobile devices and the number of passwords required of them increase [12], understanding the input of secure passwords on mobile devices is becoming increasingly important. The predictive algorithms that many users rely on for text entry on mobile touchscreen devices (like autocorrect, autocomplete, and word suggestions), are not useful—indeed, those features are disabled entirely in secure text fields—for password entry. Furthermore, the cost of errors for users differs greatly between text entry for communicative purposes (e.g., composing text messages and emails) versus text entry for authentication to a user account. In other words, the motivation for accuracy, i.e., error-avoidance, is different between tasks: while misspelled words in texts and emails can cause amusement and embarrassment, mistyped passwords can cause a user account to be locked, requiring additional steps, time, and effort to perform an account reset/unlock.

It is likely that users are sensitive to the high cost of error recovery associated specifically with password entry. Those users who are usually fast and inaccurate, relying on predictive text correction algorithms of their smartphones and tablets, may be more likely to intentionally adjust their strategy when entering passwords. In contrast, users who are generally slow and accurate may not need to adjust their speed-accuracy tradeoff function when transitioning between normal text entry and password typing tasks. Regardless, user text entry proficiency should decrease with increasing keyboard screen depth—after all, manufacturers order their screens based on frequency of use. For the more common punctuation symbols, such as a period on iOS[1] devices, it is not even necessary for a user to change screen depth. Double-tapping the space bar will automatically insert a period at the end of a sentence; this is a default keyboard setting on iOS devices, as is automatic capitalization of words following a period. Both of these conveniences are overall quite helpful during normal text entry, but again, cannot be used during password entry.

Visibility of numbers and special characters differs significantly between traditional physical keyboards in the desktop environment and onscreen keyboards on

[1] Disclaimer: Any mention of commercial products or reference to commercial organizations is for information only; it does not imply recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that the products mentioned are necessarily the best available for the purpose.

mobile devices. On the former, they are always present and visible, whereas on mobile devices, shifting between multiple screens with different character keyboards is necessary to find these numbers and special characters. While a few of the more common punctuation symbols are on the first screen of the iPad, which are not available from the first screen of the iPhone, all numbers and the majority of special characters are on different screens regardless of device (Figure 1).

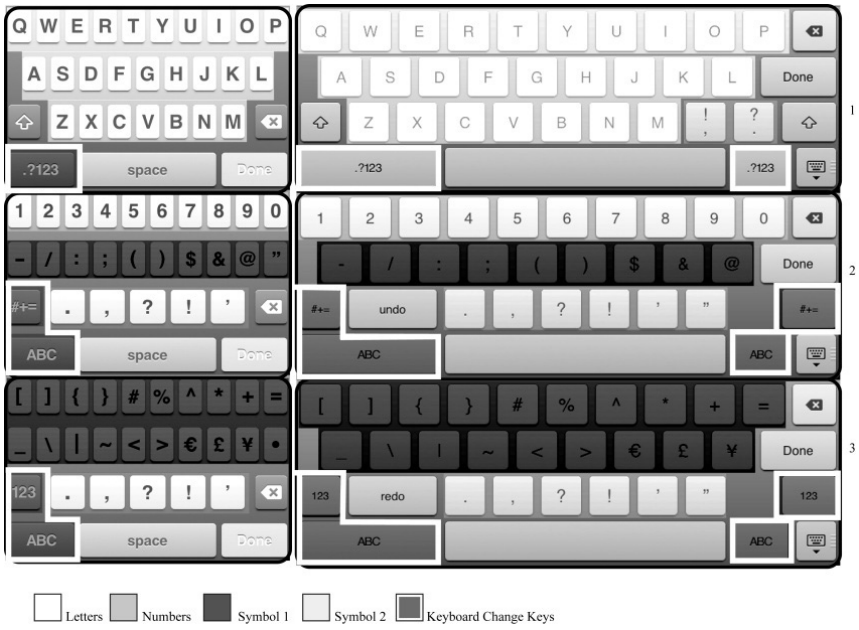

Fig. 1. The three keyboard screen depths (*top to bottom)* for iPhone (*left*) and iPad (*right*). Note that what appears on the Keyboard Change Keys differs by screen depth. Not to scale.

Multiple keyboard screens have significant perceptual-motor and cognitive implications for users. Not only can this double or triple the number of user motor actions (taps) required to input the same symbol with an onscreen keyboard compared to a physical keyboard, but multiple screen depths also carry significant cognitive overhead as well: now users must keep track of a character's position within a password, its spatial location on the visible keyboard, and its relative screen depth location. This becomes even more complicated if the current character is available on multiple screens. To investigate such issues, it is critical to have some record of user shift actions and keyboard changes, as an abundance of extraneous keyboard changes during password entry may indicate that users are indeed unfamiliar with the screen depth of special symbols, and are "losing their place" while visually searching different screens for them.

3 Experiment

3.1 Method

The current work was based heavily on a previously conducted study [11], which examined memorability of ten randomly generated, password-like character strings in the desktop computing environment; unless otherwise noted, current methodology was identical to that of [11]. We replicated this work in two studies with mobile touch screen devices, using a smartphone and tablet, respectively. To facilitate more direct comparisons, mobile device was used as a between-subjects variable in the following consolidated analyses (prerequisite random sampling assumptions were met).

Participants. Participants were recruited from the larger Washington, DC, USA metropolitan area, and were paid $75 for their participation. Participants were fairly diverse in terms of education, ethnicity, and income. A total of 165 people participated. Of these, seven did not make it at least halfway through the study session; their data were not included in any of the following analyses. The remaining 158 participants ranged in age from 19 to 66, with a mean age of 33.2 years (*SD* = 11). Ninety participants were female, and 68 were male. All were familiar with onscreen keyboards (Table 1), with 75% of participants reporting using the onscreen keyboard multiple times per day.

Table 1. Self-reported onscreen keyboard frequency of use

Frequency of use	N	%
Monthly or less	7	4.4
Weekly	18	11.4
Once a day	14	8.9
Multiple times a day	118	74.7
No report	1	0.6

Design. The experiment was a 10 (strings) x 10 (entry repetitions) x 2 (device) x 2 (age) Mixed Factorial design. All participants typed all 10 strings 10 times each (within-subjects factors of string and entry repetitions, respectively), for a total of 100 string entries per participant. Each participant used either a smartphone (iPhone 4S) or a tablet (iPad 3) to enter the strings; assignment to this between-subjects factor (device) was random. Age was the second between-subjects factor; participants were assigned to the younger or older age group based on whether their age was below or above the median age of 28 years. Eight participants were exactly the median age; they were randomly assigned to the older or younger age category.

Materials. Strings were those used in the desktop study [11], presented in Helvetica font. Participants received the strings in the same randomly determined order shown in Table 2. The data collection application was developed in-house for iOS 6.1.

Table 2. Strings by presentation order and length

Order	String	Length
1	5c2'Qe	6
2	m#o)fp^2aRf207	14
3	m3)61fHw	8
4	d51)u4;X3wrf	12
5	p4d46*3TxY	10
6	q80<U/C2mv	10
7	6n04%Ei'Hm3V	12
8	4i_55fQ$2Mnh30	14
9	3.bH1o	6
10	a7t?C2#	7[2]

Procedure. As described in [11], participants saw a series of three screens (Figure 2), corresponding to memorize/practice at will, verify correctly once, enter string 10 times. After completing this sequence for all 10 strings, a surprise recall test followed. Instructions on the surprise recall screen simply asked participants to type as many of the character strings as they could remember (they could be entered in any order). Aside from the instructions, the recall and entry screens were nearly identical, therefore the recall screen is not shown in Figure 2. Typed text was visible during memorize and verify phases, and masked with default iOS bullets during entry and recall phases.

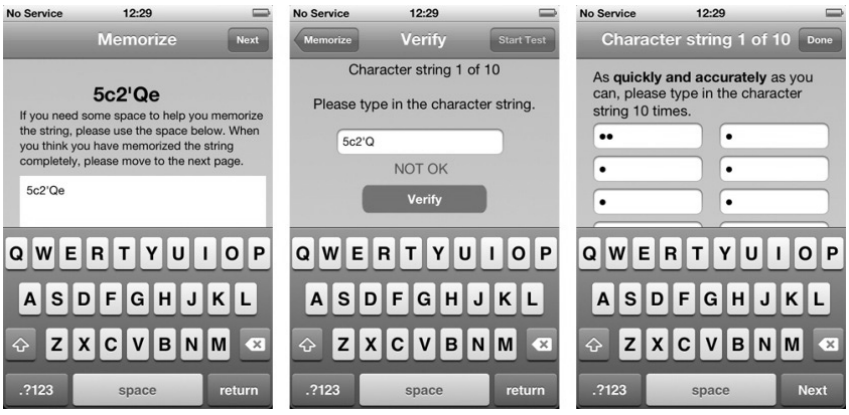

Fig. 2. Screenshots of memorize, verify, and entry screens for iPhone

[2] Note that in [11], string 10 was of length 8 rather than 7; it was preceded by the letter "u". Due to a software configuration file change, the leading "u" was omitted in the current study.

3.2 Results

Entry Times. To examine predicted effects of device and age on mean per-string text entry times, a repeated measures ANOVA was run on string by device by age. For each string, the individual 10 entry repetition times were averaged to create mean entry time measures. Observations more than three interquartile ranges (IQRs) below the 25th or above the 75th percentiles of the per-string mean entry time distributions were considered outliers and excluded from the analysis; a total of 19 participants were excluded[3] this way (seven smartphone and 12 tablet) for the following entry times analyses. Despite being unable to include data from these participants, several significant and interesting interactions and main effects were found. While longer strings in general took longer to enter, the pattern of results did not exactly follow those predicted solely by string length, nor by number of keystrokes (Table 3). Compare the two strings of length 14: String 2 requires one fewer keystrokes than String 8, yet its mean entry time is over three seconds slower, perhaps because it requires one extra screen depth change. However, screen depth changes alone do not fully predict times. String 3 requires two fewer screen depth changes than Strings 9 and 1, but is slower than both of them; while number of keystrokes is equivalent, string 3 is longer in length, so it contains more characters for a person to recall. Clearly, a combination of factors account for entry times, with screen depth changes a factor unique to mobile devices..

Not surprisingly, older participants were overall somewhat slower than were younger participants. While these timing differences were negligible and consistent for the easier strings (strings 1, 3, 9, and 10), they were more pronounced for the more difficult strings (2, and 4 through 8). Overall, tablet string entry times were faster than the corresponding smartphone times. Mean entry times between devices did not differ significantly for the hardest string (string 2), suggesting that screen switches may be equally cognitively disruptive regardless of device, and/or that the visual search time for special symbols on the second and third screen depths is problematic regardless of device. Mean entry times were also similar between devices for the easiest strings (strings 1, 3, 9, and 10). The main effect of string on mean entry times (Fig. 3) was significant ($F(5.04, 594.69) = 468.99$, $MSE = 7392.49$, $p < .001$, $\eta_p^2 = .80$, Greenhouse-Geisser adjustment), as was the interaction between string and device (Fig. 4) ($F(5.04, 594.69) = 2.46$, $MSE = 38.77$, $p = .03$, $\eta_p^2 = .02$, Greenhouse-Geisser adjustment). The interaction between string and age was also reliable (Fig. 5) ($F(5.04, 594.69) = 2.23$, $MSE = 35.15$, $p = .05$, $\eta_p^2 = .02$, Greenhouse-Geisser adjustment). The main effect of device was significant ($F(1, 118) = 11.01$, $p = .001$, $\eta_p^2 = .09$), as was the main effect of age ($F(1, 118) = 15.16$, $p < .001$, $\eta_p^2 = .11$).

[3] While there were several alternative outlier replacement methods we could have used (e.g., replace the observation with that participant's mean, with that string's entry mean, or with the grand mean), we chose to consistently exclude participants instead, as it was unclear whether any alternative was better justified given the large variability seen in our data.

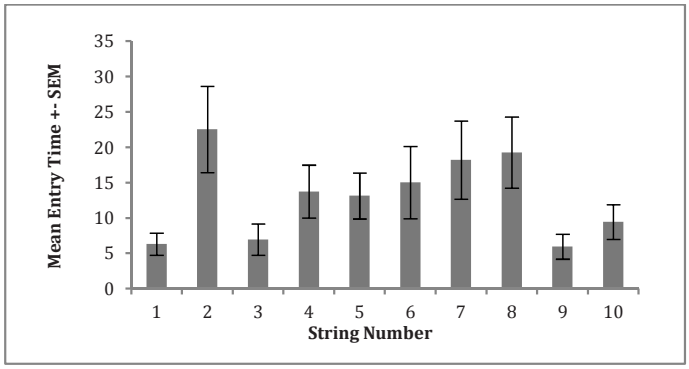

Fig. 3. Significant main effect of string on mean text entry times (seconds)

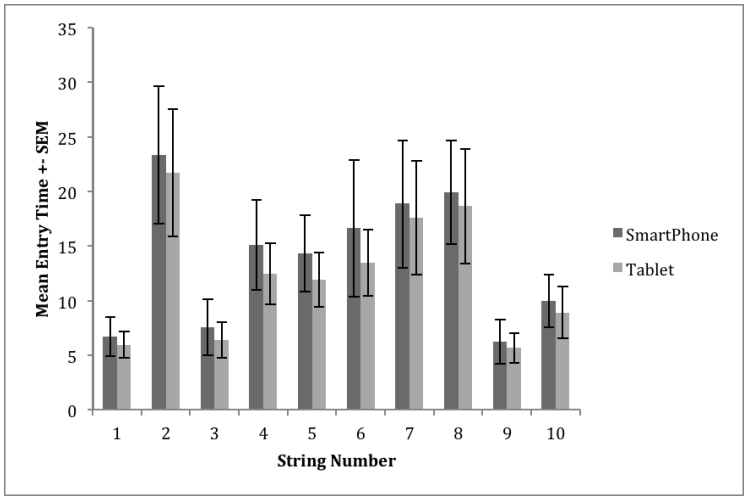

Fig. 4. Significant interaction of string by device on mean text entry times (seconds)

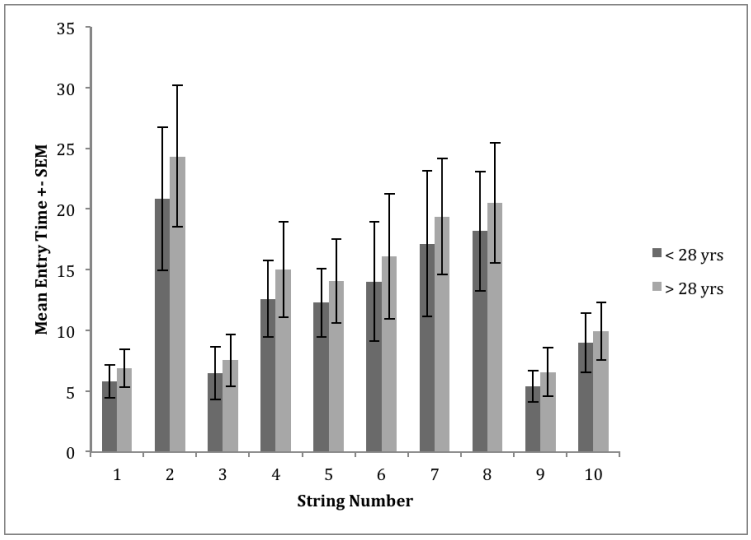

Fig. 5. Significant interaction of string by age on mean text entry times (seconds)

Table 3. Per-string lengths, keystrokes, shifts, and screen depth changes (taps on keyboard change keys), presented from shortest to longest mean entry time (seconds)

Order	String	Mean Entry Time	Length	Key-strokes	Shifts	Screen depth changes
9	3.bH1o	5.97	6	11	1	4
1	5c2'Qe	6.32	6	11	1	4
3	m3)61fHw	6.98	8	11	1	2
10	a7t?C2#	9.45	7	14, 13*	1, 2*	6, 4*
5	p4d46*3TxY	13.13	10	18	2	6
4	d51)u4;X3wrf	13.75	12	19	1	6
6	q80<U/C2mv	15.02	10	19	2	7
7	6n04%Ei'Hm3V	18.20	12	24	3	9
8	4i_55fQ$2Mnh30	19.28	14	25	2	9
2	m#o)fp^2aRf207	22.52	14	24	1	10

*(iPhone, iPad)

Memorize and Verify Times. Since participants were free to spend as much or as little time on the memorization phase as they pleased, and could revisit the memorize screen at will from the verify screen, the number of visits to both the memorize and verify screens differed widely by participant. Therefore, for these two measures we report total times rather than mean times. Total memorize time and total verify time represent summations of all time (across multiple visits) a participant spent on each screen, respectively. As the patterns of results for total memorize and total verify times were similar to those reported for mean entry times above, we do not present additional figures for significant results in this section. In sharp contrast to the number of extreme entry time observations reported above, only three (two iPhone, one iPad) total memorize time observations were more than three interquartile ranges (IQRs) below the 25th or above the 75th percentile of the memorize total time distribution. These observations were considered outliers and excluded from the following memorize time analyses. The main effect of string on total memorize times was significant, ($F(3.75, 502.07) = 219.05$, $MSE = 1229047.63$, $p < .001$, $\eta_p^2 = .62$, Greenhouse-Geisser adjustment), as was the interaction between string and age, ($F(3.75, 502.07) = 9.30$, $MSE = 52184.44$, $p < .001$, $\eta_p^2 = .07$, Greenhouse-Geisser adjustment). The main effects of device, ($F(1, 134) = 13.48$, $MSE = 282428.253$, $p < .001$, $\eta_p^2 = .09$) and age, ($F(1, 134) = 10.47$, $MSE = 219295.23$, $p = .002$, $\eta_p^2 = .07$) were again significant.

As with entry times, there were numerous extreme observations when examining total verification time. Using the same outlier definition as above, a total of 30 participants (13 smartphone and 17 tablet) were excluded from the following analysis. The average number of failed verify attempts on the iPhone was 7.69 with a standard deviation of 8.65, for the iPad it was 3.65 with a standard deviation of 6.94. The main effect of string on total verify times was significant, ($F(3.33, 356.62) = 63.71$, $MSE = 65166.53$, $p < .001$, $\eta_p^2 = .37$, Greenhouse-Geisser adjustment), as was the interaction

between string and device, ($F(3.33, 356.62) = 4.11$, $MSE = 4198.76$, $p = .01$, $\eta_p^2 = .04$, Greenhouse-Geisser adjustment). The main effects of device, ($F(1, 107) = 13.10$, $MSE = 18689.25$, $p < .001$, $\eta_p^2 = .11$) and age ($F(1, 107) = 4.47$, $MSE = 6373.00$, $p = .037$, $\eta_p^2 = .040$) on total verify times were again significant.

Entry Errors. Each string that was typed in the entry phase was analyzed based on the errors it contained at the time of final submission. Based on the common types of errors considered in text entry experiments and the frequency of certain errors types found in this experiment, the following subcategories were created. Extra Character errors occurred when duplicate or additional characters were entered into the field. Missing Character errors occurred when characters were omitted from entry. There were four types of substitution errors: substitution of the correct character with a Wrong Character; with an Incorrectly Shifted character; with an Adjacent Key character (with a character adjacent to it on the keyboard, for example, Q's adjacent keys are A and W); and substituting the number zero for the letter "o" and vice versa (while this could also be considered a Wrong Character error, its high frequency of occurrence warranted giving it a separate category. There were two types of transposition errors: transposition of characters next to one another in the string and characters typed in the wrong place in the string, referred to as Transposition and Misplaced Character, respectively.

Both the frequency and nature of errors varied greatly by device. With the smartphone, there were a total of 2100 errors made, as compared to 1289 errors with the tablet. Most interestingly, the percentage of adjacent key errors was much higher for the smartphone than the tablet (Fig. 6). The onscreen keys are much smaller targets on an iPhone than an iPad overall, and are particularly problematic for the iPhone portrait orientation. Given that participants were forced to use the devices in portrait rather than landscape orientation, the difference in adjacent characters would be expected.

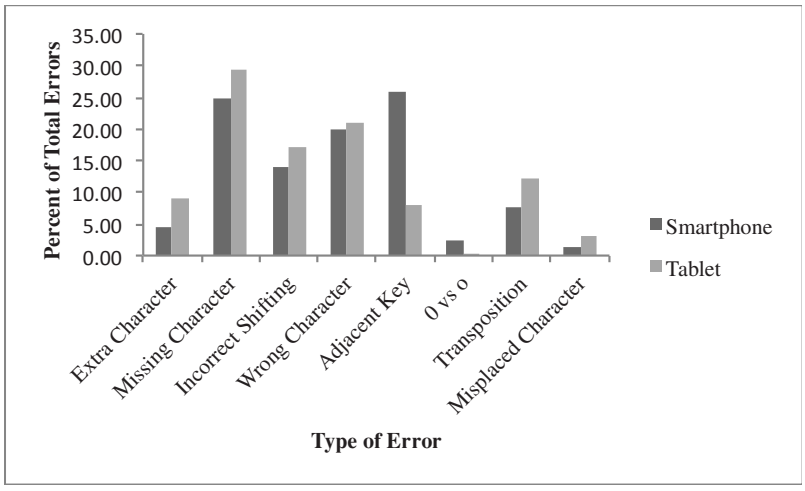

Fig. 6. Percentages of entry errors by error category and device

Surprise Recall. A string was considered correctly (fully) recalled if it exactly matched one of the target strings. Forty-six participants did not recall any of the strings correctly. Most participants were able to recall one or two strings, with one participant able to recall seven strings (Table 4). The most frequently recalled string was the last string memorized, followed by the second-to-last string memorized for each study (Table 5).

Table 4. The frequency of fully-recalled strings

String	Total Times Recalled
a7t?C2#	104
3.bH1o	44
4i_55fQ$2Mnh30	28
q80<U/C2mv	7
6n04%Ei'Hm3V	4
d51)u4;X3wrf	3
p4d46*3TxY	3
m#o)fp^2aRf207	2
5c2'Qe	0
m3)61fHw	0

Table 5. Number of strings fully-recalled by participants during surprise recall task

Number of Strings Recalled	Number of Participants
0	46
1	56
2	37
3	11
4	5
5	2
7	1

Note that in [11], text was visible during the recall phase, whereas in our experiment, text was masked during surprise recall. While this may account for some differences in recall performance between studies, it is more likely that device played a much more significant role.

4 Discussion

In general, main effects are not typically as interesting as interactions, but in this case the underlying explanation behind the main effect of string on both errors and times is at the core of our findings: strings requiring a number of mobile screen depth changes have disproportionately large effects across a variety of dependent measures. They are physically more difficult to type and error-prone, especially for adjacent key characters in the smartphone portrait orientation. Yet mobile devices affect password entry for reasons beyond simply smaller key sizes; these devices can place significantly more demands on working memory for users. Screen depth changes are like mini task interruptions that seem to incur timing costs beyond simply the additional keystrokes (i.e., taps on keyboard change keys) required. People are sensitive to the interruption cost of screen depth changes. As one participant noted, "My brain can't focus on memorizing it. It has to focus on find the right key board. [sic] Now that is a challenge." Clearly, password entry on mobile devices is challenging both cognitively and motorically. We argue that there are platform-dependent cognitive components associated with the interruptive nature of back-and-forth navigation and searching between mobile screen depths. This suggests that simply porting password requirements from one platform to another (i.e., desktop to mobile) without considering device constraints may be unwise.

5 Limitations and Future Work

In the future, we hope to better disentangle typing from memory errors. Our current error analysis has the limitation that it is not utilizing all the data available in the input stream but instead is focused on classified errors that were left unfixed in the text upon submission. A mobile transcription typing experiment that uses password-like text as stimuli, with a full input stream error analysis, would further aid in determining the nature and frequency of typing errors for complex passwords. Ultimately, we need these data to inform and validate predictive, computational cognitive models of password entry on mobile touch screens. Such models can help more objectively evaluate the benefits of additional security requirements against the drawbacks of more onerous passwords for users. To examine the effects of changing password policies over time and across devices, the research community needs fine-grained, baseline human performance data to which we can compare emerging and future technologies and text entry methods.

Working with newer mobile technologies presents interesting challenges that must be addressed in future work. For example, one challenge with using iOS devices is that the only keyboard change event reported by the native iOS keyboard is the show/hide keyboard event; keyboards taps that do not result in changing text are not reported by the OS. This means that taps on the keyboard change keys themselves are not reported, as they do not cause any visible evidence in the text entered. Simply examining interkey intervals in the entered text would not completely address this fundamental piece of password entry that is specific to mobile devices, i.e., that

complex passwords force users to deal with numerous screen depth changes. While one can infer that a user had to have tapped on a keyboard change key in order to enter a particular character given the preceding character in the input stream, one would not know how many times the keyboard changed, nor the associated keystroke latencies for each event. This is important future work, as screen depth changes are fundamental differences between password entry with onscreen versus physical keyboards.

Acknowledgements. The authors gratefully acknowledge Clayton Stanley at Rice University. This work was funded by the Comprehensive National Cybersecurity Initiative (CNCI).

References

1. Honan, M.: Kill the password: Why a string of characters can't protect us anymore. Wired (2012)
2. United States Department of Homeland Security: United States Computer Emergency Readiness Team (US-CERT). Security tip (ST04-002): Choosing and protecting passwords (2009), retrieved from website, http://www.us-cert.gov/cas/tips/ST04-002.htm
3. Stanton, B.C., Greene, K.K.: Character strings, memory and passwords: What a recall study can tell us. In: Tryfonas, T., Askoxylakis, I. (eds.) HAS 2014. LNCS, vol. 8533, pp. 195–206. Springer, Heidelberg (2014)
4. Arif, A.S., Lopez, M.H., Stuerzlinger, W.: Two new mobile touchscreen text entry techniques. In: Poster at the 36th Graphics Interface Conference, pp. 22–23 (2010)
5. Nicolau, H., Jorge, J.: Elderly text-entry performance on touchscreens. In: Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility. ACM, Boulder (2012)
6. Nicolau, H., Jorge, J.: Touch typing using thumbs: understanding the effect of mobility and hand posture. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2683–2686 (2012)
7. Castellucci, S.J., MacKenzie, I.S.: Gathering text entry metrics on android devices. In: CHI 2011 Extended Abstracts on Human Factors in Computing Systems, pp. 1507–1512 (2011)
8. Parisod, A., Kehoe, A., Corcoran, F.: Considering appropriate metrics for light text entry. In: Fourth Irish Human Computer Interaction Conference, Dublin City University (2010)
9. Sears, A., Zha, Y.: Data entry for mobile devices using soft keyboards: Understanding the effects of keyboard size and user tasks. International Journal of Human-Computer Interaction 16(2), 163–184 (2003)
10. Allen, J.M., McFarlin, L.A., Green, T.: An in-depth look into the text entry user experience on the iPhone. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 52(5), pp. 508–512 (2008)
11. Salthouse, T.: Effects of age and skill in typing. Journal of Experimental Psychology 113(3), 345–371 (1984)
12. Florencio, D., Herley, C.: A large-scale study of web password habits. In: Proceedings of the 16th international conference on World Wide Web 2007, pp. 657–666 (2007)