Character Strings, Memory and Passwords: What a Recall Study Can Tell Us*

Brian C. Stanton and Kristen K. Greene

National Institute of Standards and Technology,
100 Bureau Dr, Gaithersburg, MD, USA
{brian.stanton,kristen.greene}@nist.gov

Abstract. Many users must authenticate to multiple systems and applications, often using different passwords, on a daily basis. At the same time, the recommendations of security experts are driving increases in the required character length and complexity of passwords. The thinking is that longer passwords will result in greater "entropy," or randomness, making them more difficult to guess. The greater complexity requires inclusion of upper- and lower-case letters, numerals, and special characters. How users interact and cope with passwords of different length and complexity is a topic of significant interest to both the computer science and cognitive science research communities.

Using experimental methodology from the behavioral sciences, we set out to answer the following question: how memorable are complex character strings of different lengths that might be used as higher-entropy passwords? In this experiment, participants were asked to memorize a series of ten different character strings and type them repeatedly into a computer program. Character string lengths varied and the random characters were made up of alphanumeric and special characters in order to mimic passwords. Not surprisingly, our findings indicate that the longer a character string is, the longer it takes for a person to recall it, and the more likely they are to make an error when trying to re-type that string. These effects are particularly pronounced for strings of eight to ten characters or longer.

Keywords: passwords, security, character strings, memory, recall.

1 Introduction

As people increasingly interact with multiple computer systems over the course of a day, they are expected to remember an ever-increasing number of passwords [5, 3]. Computer security specialists also want to increase the length of these passwords in order to increase their "entropy," or randomness, making them more difficult to guess. This means users are often forced to remember not only more passwords but longer passwords as well. Increasing password length is not the only method of increasing password entropy; another option is increasing password complexity. The

* 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. 195–206, 2014.

inclusion of upper- and lower-case letters, numerals, and special characters are often recommended for increasing password security [13]. How users interact and cope with the increasing level of complexity of passwords is an area of interest to the usable security community. This research explored the following question: how memorable are complex character strings of different lengths that might be used as higher-entropy passwords? Password memorability is a multi-faceted concept, affecting the amount of time required to initially commit the password to memory; the time to recall and type the password; and the nature and frequency of errors committed during password entry.

2 Background

In order to provide best practice recommendations for institution-wide password policies, it is critical that the usable security field better understands how various password requirements fundamentally affect human performance. The nature of the interplay between password complexity, errors, timing, and memorability should be more closely examined. It has been long remarked that longer passwords "take longer to enter, have more chance of error when being entered, and are generally more difficult to remember" [12]. It is to be expected that longer passwords should lead to longer entry times and more errors (more characters offer more chances for misremembering) but how many characters are too many? When does the burden of remembering become too much for a user and what types of errors do users make when recalling and typing passwords?

There have been many studies of remembering in general (e.g., [14]) and passwords in particular, addressing such issues as memorability, predictability and attention [6, 7, 15, 16, 17, 18]. In addition, there is a large body of literature examining the factors of skilled typing performance from 1923 [4] through the 1980s [8, 10] including a great deal of literature on the cognitive and perceptual-motor aspects of transcription typing [11]. But, comparatively little research has been done on the fundamentals of password typing. Secure passwords differ greatly from the words used in traditional transcription typing studies; the former are ideally as random as possible, whereas the latter follow orthographic rules and are easily predictable given the surrounding semantic content. Although non-word strings of random letters have been studied in previous transcription typing research (e.g., [10]), such research did not include the variety of numbers and special characters recommended for passwords. The current study is a necessary first step in addressing the fundamentals of passwords.

3 Method

3.1 Participants

Two groups of participants were tested in this study. The first group consisted of 30 participants recruited from the Washington, DC (WDC) metropolitan area in the

United States. Seven WDC participants failed to complete the test in the one-hour time allotted. The second group consisted of 45 participants recruited from the University College London (UCL) in the United Kingdom. All UCL participants completed the test in the time allotted. Ages ranged from 18 to 78 for the two groups with most of the UCL group between the ages of 18 – 27.

3.2 Instructional Materials

The participants were given the following verbal instructions:

You will be working on this computer. You will be presented with 10 character strings with varying lengths, one at a time. Your task is to memorize each string as it's presented to you on the screen. You can take as much time as you need to memorize each string. You may also practice typing the string. After you feel that you have the string memorized, you will be given the chance to verify that you have memorized the string. If you don't pass the verification, you can re-try the verification or go back and memorize the string again. If you do pass the verification, you will be asked to type the character string in ten times. After typing the string in ten times, you will move on to the next character string.

3.3 Materials and Equipment

The strings the participants were asked to memorize consisted of ten strings made up of two strings each of six, eight, ten, twelve, and fourteen character strings. The strings were randomly generated, using a software package1. Each string consisted of upper case, lower case, alphabetic, numeric, and special characters presented in the Consolas font. Strings could not begin with a capital letter, nor could they end with an exclamation mark. The strings are shown in Table 1.

By using randomly generated character strings as stand-ins for user-generated passwords, we hoped to control for effects of different levels of password meaningfulness. Since we only set out to study effects of increasing password length, we wanted to keep other factors, such as meaningfulness, constant across stimuli. Rather than making stimuli equally memorable, we wanted them to be equally unmemorable.

The strings were presented in the above random order for all participants in the WDC group. The last two strings were switched for the UCL group due to a software configuration file change. A custom software program was designed to present the strings, allow the user to enter the strings, and time the user actions.

[1] Advanced Password Generator from BinaryMark was used. Disclaimer: Any mention of commercial products is for information only; such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that these entities, materials, or equipment are necessarily the best for the purpose.

Table 1. Character string order, string, 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	ua7t?C2#	8

Both studies were conducted using a desktop PC with monitor, keyboard and mouse. The WDC study used a standard American QWERTY keyboard, while the UCL study used a standard UK QWERTY keyboard.

3.4 Data Collection Methods

The participant was verbally given the instructions quoted above then was given an informed consent form to sign. The test facilitator started the data collection program and entered the participant number. The participant was given a piece of paper with the first character string. The participant was presented with instructions on the practice screen asking them to memorize the target string (see Fig. 1. Practice Screen). When the participant felt that they had memorized the target string they moved to the next screen (see Fig 2. Verification Screen).

The second screen asked the participant to enter the memorized target string. The string had to be entered correctly in order to move to the third screen where the participant was asked to enter the memorized string ten times (see Fig 3. Entry Screen). If the entered string failed the verification, the user had the opportunity to go back to the practice screen or they could try and enter the string again.

This procedure was repeated for the ten strings. After all ten strings had been tested, the program gave them a surprise recall test to see how many of the ten strings they remembered. During the surprise recall test, if participants asked, they were informed that they didn't have to type the ten strings in the sequence in which they were presented. Typed text was visible during practice and verification (Figs. 1 and 2, respectively), masked with asterisks during entry (Fig. 3), then visible during surprise recall (not shown given its high similarity to the entry screen, Fig. 3).

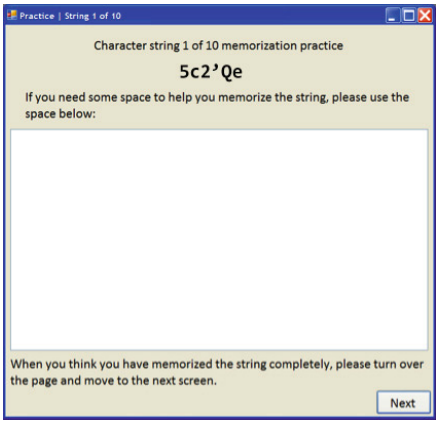

Fig. 1. Practice Screen

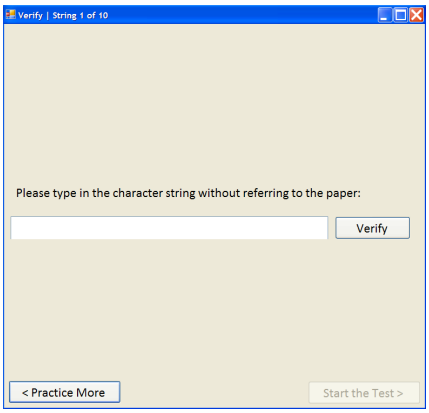

Fig. 2. Verification Screen

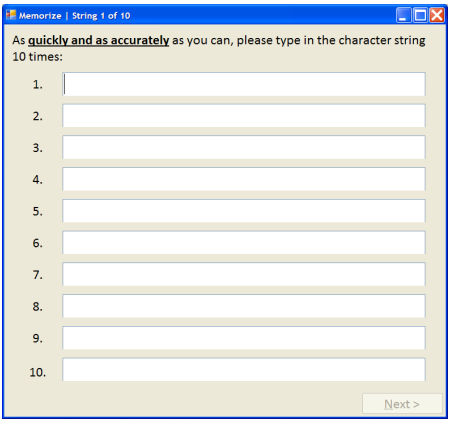

Fig. 3. Entry Screen

4 Results

The results were analyzed for the amount and types of errors made during the individual string entry, the amount of time for each individual string entry and the number of strings correctly recalled during the surprise recall task.

4.1 Errors for Entry Tasks

An entered string was in error if it did not exactly match the target string. As long as the entered string contained at least one deviation from the target string, it was deemed to be in error. Fig. 4 and Fig. 5 show the median errors per character string length. The UCL participants made fewer median errors overall than did the WDC group when the character string length reached 10 or greater. Both groups had increases in the variability of error counts as the string length increased.

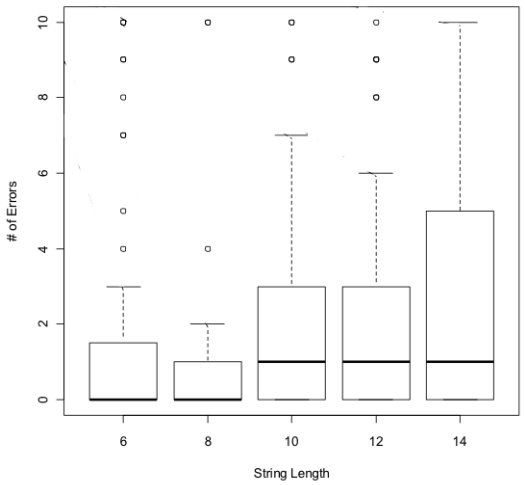

Fig. 4. WDC Errors

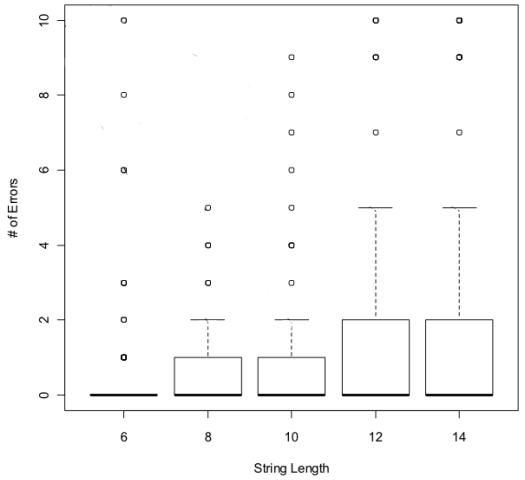

Fig. 5. UCL Errors

4.2 Types of Entry Errors

Each entered string was analyzed as to the type of error or errors it contained. The types of errors made were as follows:

- Extra character
- Missing character
- Incorrect capitalization (shifting)

- Wrong character
- Character typed was adjacent on the keyboard to the target character
- Zero instead of an "O"
- Transposition of characters next to one another in the string
- Character was in the wrong place within the string (misplaced character).

Typing a zero rather than an "O" occurred often enough as to deserve its own category rather than being grouped into the "wrong character" category. The WDC group made 471 errors and the UCL group made 556 for a total of 1,027 errors.

Table 2. Types of entry errors made

Type of error	WDC Percentages	UCL Percentages	Total Percentages
Extra character	7%	7%	7%
Missing character	25%	10%	17%
Incorrect capitalization	38%	51%	45%
Wrong character	6%	12%	9%
Adjacent key	8%	10%	9%
Zero instead of an "O"	3%	3%	3%
Transposition of characters	10%	6%	8%
Wrong place within the string	3%	1%	2%

4.3 Task Times

The time for a task was calculated from the time the practice screen was first presented until the tenth recalled string was entered on the entry screen. As Fig 6. shows, the average time taken to complete the task increases as the character string

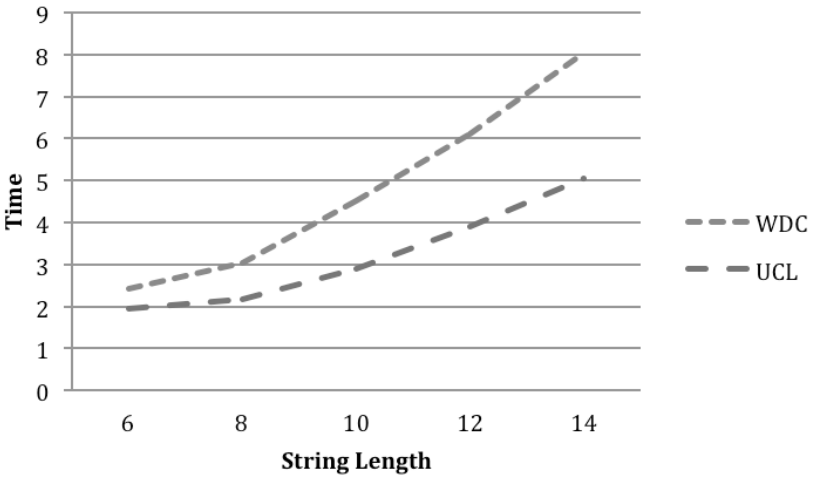

Fig. 6. Average task time (minutes) by character string length

length increases. The figure also shows that the rate of increase increases after the eight character string for both groups, with the WDC group taking somewhat longer overall. Table 3 and Table 4 show the minimum, maximum, mean, and standard deviation (SD) task times for the two groups.

Table 3. WDC task times (minutes) by string length

Length	N	*Min.*	*Max.*	Mean	*SD*
6	24	1.05	4.24	2.4691	0.82538
8	24	1.27	5.17	3.0723	0.99511
10	24	1.65	7.65	4.5305	1.50970
12	24	2.41	9.67	6.1832	1.96270
14	24	3.14	12.83	8.0683	2.55134

Table 4. UCL task times (minutes) by string length

Length	N	Min.	Max.	Mean	SD
6	45	1.43	3.98	2.3410	0.50364
8	45	1.05	2.92	1.7874	0.38765
10	45	1.61	5.79	2.9255	0.86637
12	45	1.98	6.84	3.9296	1.11102
14	45	2.61	10.62	5.0954	1.67292

4.4 Number of Surprise Strings Recalled

Roughly one half of the participants could only remember one string with only one person recalling the maximum number of four strings (see Table 5).

Table 5. Number of surprise strings recalled

Strings Recalled	Number of WDC Participants	Number of UCL Participants
0	5	3
1	12	21
2	5	17
3	2	3
4	0	1

The most recalled string during the surprise recall task was the last string memorized followed by the second to the last memorized string for each group (see Table 6).

Table 6. Surprise strings recalled.

String	Number of times recalled	
	WDC	UCL
ua7t?C2#	17	21
3.bH1o	7	33
q80<U/C2mv	2	2
4i_55fQ$2Mnh30	1	10
6n04%Ei'Hm3V	1	1
m#o)fp^2aRf207	0	1

5 Discussion

In his 1956 paper on human information processing, Miller proposed that human short-term memory could only retain seven plus or minus two items [9]. If we surmise that our participants are working from short-term memory only (or what Baddeley and Hitch called "working memory" [1]), then the results of our study seem to bear out Miller's assertion. This supposition is supported by the final surprise recall results, which show that most participants could only correctly recall the most recent strings they had worked with (see Table 6). We expect that the participants may have been able to recall more strings if the strings had been committed to long-term memory.

As it is likely that the character strings did not go into the participants' long-term memories, we would expect recall success to decrease around the eight- and ten-character string lengths, since that is the point where the number of items to recall would begin exceeding the "seven plus or minus two" range. We found these changes for both timing and errors. Given that, we were not surprised by the finding that the longer the character string was, the more time it took for participants to complete the tasks. What is interesting is that the slope of the timing line increased around the eight-character string length for both the WDC and UCL groups, even though UCL participants were faster overall (see Fig. 6). This would suggest that there is added work involved when the string length exceeds eight characters (as is predicted by [9]). The finding that the UCL participants were faster at completing the tasks may potentially be explained by the fact that they were sampled from a younger participant pool (UCL college students), and may therefore have had better typing skills and/or working memory capacities than the (on average) older WDC participants, who were sampled from the larger Washington DC metropolitan area.

The median number of errors also increased around the eight- to ten-character string lengths. This trend was more visible in the WDC data (see Fig. 4), where the median number of errors increased from zero to one between the eight- and ten-character strings. The variability of the error counts increased at the same point. Even though the median number of errors remained the same for the UCL participants (see Fig. 5), they experienced increased variability of errors around the ten- to twelve-character strings. As with the difference in entry times between the WDC and UCL groups, the difference in the "error variability threshold" may potentially be explained

by the younger UCL participants having greater memory capacity and/or better general typing skills. To test these potential explanations, future studies should collect data on whether participants are touch typists and measure their Words Per Minute (WPM) for typing prose passages, in order to account for individual differences in general typing ability. It will also be necessary to capture more granular data on participants' ages; while we know the age ranges from which each group in the current study were recruited, we unfortunately did not have access to ages at the individual participant level. Future studies would also benefit from administering a standardized battery of cognitive ability tests to quantify effects of individual differences in memory capacity.

One of the more interesting findings was the type of errors made. In both participant groups, the largest percentage of errors were capitalization (shifting) errors (see Table 5). Many special characters also require a shift action (e.g., "8" must be shifted to "*"), so these errors are particularly important given the increasing use of special characters in password policies.

6 Conclusions and Future Work

Since capturing real-world password typing data poses significant privacy and security concerns, we instead gathered human performance data in a controlled laboratory experiment using randomly generated, password-like character strings. Admittedly, having randomly generated character strings represent passwords is somewhat artificial, since people often (but not always) create and use passwords that have some meaning for them. Still, we feel that some general recommendations can be derived from our results.

First, the trend towards ever-increasing password lengths is likely to be problematic for users. Our results indicate that the longer a character string is, the longer it takes a person to memorize, recall, and enter it. Longer strings also increase the probability of errors.

Secondly, the trend of requiring special characters and capital letters should be weighed against the increased likelihood of errors, especially for those systems that limit the number of password attempts before lockout. It is possible that longer passwords with more special characters and capital letters may require more attempts to enter them correctly than passwords with fewer (or no) special characters and capital letters. This means that some organizations may need to consider changing the typical "three strikes, you're out" policy for password attempts.

With regards to conducting further research, a natural next step would be to replicate this experiment but have the participants choose their own passwords instead of issuing them random character strings. Participants would likely find chosen (as opposed to assigned) passwords more meaningful and therefore easier to remember. Although challenging from a security and privacy perspective, it would nonetheless be interesting to see whether the "seven plus or minus two" rule would still be in effect in such a case. Another extension of interest would be replicating this experiment on different platforms, such as smartphones. How would working on a

different platform affect the input of passwords? Are people using alternative platforms faster or slower when inputting password-like character strings? Do they make more or fewer errors? Are individual differences such as cognitive ability, touch-typing ability, and age more pronounced in the mobile computing environment? These are all questions that bear further investigation.

This study contributes to the usable security community by presenting much-needed human performance data that are difficult to obtain in the real world. This is the first in a series of planned studies exploring effects of password requirements across platforms, starting with the traditional desktop environment and moving on to mobile devices. Only by understanding the fundamental characteristics of password typing may we hope to predict how well users will be able to comply with proposed password policy changes.

Acknowledgments. We would like to thank Dr. Angela Sasse and her colleagues at the University College London for the UCL data collection.

References

1. Baddeley, A.D., Hitch, G.: Working memory. In: Bower, G. (ed.) Recent Advances in Learning and Motivation, vol. 8, pp. 47–90. Academic Press, New York (1974)
2. Chiasson, S., Forget, A., Stobert, E., Van Oorschot, P., Biddle, R.: Multiple password interference in text passwords and click-based graphical passwords. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, pp. 500–511 (2009)
3. Choong, Y., Theofanos, M., Liu, H.: A Large-Scale Survey of Employees' Password Behaviors. Manuscript submitted for publication (manuscript in preparation, 2014)
4. Coover, J.E.: A method of teaching typewriting based upon a psychological analysis of expert typing. National Education Association 61, 561–567 (1923)
5. Florencio, D., Herley, C.: A large-scale study of web password habits. In: WWW 2007, Banff, Canada. ACM Press (2007)
6. Forget, A., Biddle, R.: Memorability of persuasive passwords. In: CHI 2008 Extended Abstracts on Human Factors in Computing Systems, pp. 3759–3764 (2008)
7. Gehringer, E.F.: Choosing passwords: Security and human factors. In: International Symposium on Technology and Society (ISTAS 2002), pp. 369–373 (2002)
8. Gentner, D.: Skilled finger movements in typing. Center for Information Processing, University of California, San Diego. CHIP Report 104 (1981)
9. Miller, G.A.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review 63(2), 81–97 (1956), doi:10.1037/h0043158
10. Salthouse, T.: Effects of age and skill in typing. Journal of Experimental Psychology 113(3), 345–371 (1984)
11. Salthouse, T.: Perceptual, cognitive, and motoric aspects of transcription typing. Psychological Bulletin 99(3), 303–319 (1986)
12. United States Department of Commerce, National Institute of Standards and Technology (NIST), Password usage (FIPS PUB 112) (1985), `http://www.itl.nist.gov/fipspubs/fip112.htm` (retrieved)

13. United States Department of Homeland Security, United States Computer Emergency Readiness Team (US-CERT), Security tip (ST04-002): Choosing and protecting passwords (2009), http://www.us-cert.gov/cas/tips/ST04-002.html (retrieved)
14. Unsworth, N., Engle, R.W.: The foundations of remembering: Essays in honor of Henry L. Roedgier III, pp. 241–258. Psychology Press, New York (2007)
15. Vu, K., Bhargav-Spantzel, A., Proctor, R.: Imposing password restrictions for multiple accounts: Impact on generation and recall of passwords. In: HFES 47th Annual Meeting, pp. 1331–1335 (2003)
16. Vu, K., Cook, J., Bhargav-Spantzel, A., Proctor, R.W.: Short- and long-term retention of passwords generated by first-letter and entire-word mnemonic methods. In: Proceedings of the 5th Annual Security Conference, Las Vegas, NV (2006)
17. Vu, K., Proctor, R., Bhargav-Spantzel, A., Tai, B., Cook, J., Schultz, E.: Improving password security and memorability to protect personal and organizational information. International Journal of Human-Computer Studies 65, 744–757 (2006)
18. Yan, J., Blackwell, A., Anderson, R., Grant, A.: Password memorability and security: Empirical results. IEEE Security & Privacy 2(5), 25–31 (2004)