Experimental Design and Stimuli Generation Challenges in Password Memorability Research

ABSTRACT

Background. Our focus is on password policies that are much more stringent than the password requirements for the personal accounts contained in most leaked databases. As we cannot use such databases for ethical and legal reasons, it was necessary to generate our own stimuli instead. Therefore, we used randomly generated character strings in place of passwords in the two studies reported here.

Aim. We set out to answer these questions: 1) how memorable are complex character strings of different lengths that might be used as higher-entropy passwords? 2) How do we define and measure password memorability?

Method. 31 participants from the Washington DC metropolitan area and 45 participants from the University College London were asked to memorize 10 character strings, one at a time. The strings consisted of two strings each of six, eight, ten, twelve, and fourteen upper, lower, alphabetic, numeric, and special characters. Memorization consisted of unlimited practice, a verification, and then typing the string 10 times. After the tenth string was entered, a surprise recall task was administered.

Results. While we fully expected that longer strings would be more error-prone (they were) and take longer to type (they did), we had not anticipated the observed non-linear increase in times. Although the relationship between length and timing was monotonically increasing, it was not strictly linear, as may have previously been assumed. Furthermore, it is unclear whether the observed timing effects were purely due to manipulating string length.

Conclusions. While we used randomly-generated stimuli, this alone failed to control for several confounding factors, such as meaningfulness. Additionally, the frequency and difficulty of special symbols was unevenly distributed across the different string lengths we studied; purely by chance, shorter strings contained easier symbols. Due to this confound, we were unable to fully disambiguate effects of string length relative to those of special symbols. Random-generation alone is insufficient to control for all potentially confounding factors.

Categories and Subject Descriptors

D.4.6 [**Security and Protection**]: Authentication; H.1.2 [User/Machine Systems]: Human factors.

General Terms

Measurement, Performance, Design, Experimentation, Security, Human Factors, Theory.

Conference'10, Month 1–2, 2010, City, State, Country.

Keywords

Human performance; passwords; password complexity; password composition; password entropy; password length; password memorability; password policy; password rules; random character strings; recall; short-term memory; system-assigned passwords; text entry; typing; usable security; working memory.

1. INTRODUCTION

For what was intended to be a relatively simple behavioral study and its replication—each examining effects of increasing length requirements on password-like character strings—we were faced with a number of technical and methodological challenges. Following the introduction and background sections, this paper includes the results of two studies, surrounded by in-depth discussion of the associated methodological decisions, mishaps, and insights. The most significant methodological challenges and decisions are detailed in the Approach section, while the ramifications and shortcomings of our methodology are revisited in the final discussion, limitations, and conclusions sections. For each research stage—experimental design and stimuli generation, programming implementation, pilot testing, data collection, analysis, and publication—we describe the methodological lessons learned. We pay special attention to the pilot testing phase, as those pilot data heavily informed several significant experimental design decisions and stimuli changes prior to the ensuing two studies.

Before beginning our password memorability studies—even before pilot testing—we first had to determine whether we would generate our own stimuli, or whether there were pre-existing password databases we could use. Although databases of leaked (e.g., the Gawker dataset) and stolen "real-world" passwords certainly exist, there are both ethical and legal concerns regarding their use; while some researchers may have ways of addressing those issues and using leaked databases responsibly, as government employees we are prohibited from doing so. Even if the use of such databases were ethically and legally permissible for us, the passwords they contain would not necessarily address our research goals. There are two reasons for this: 1) Our focus is on governmental password policies that are much more stringent than the password requirements for the personal accounts contained in most leaked databases, and 2) Leaked or not, mere password lists would not provide the fine-grained data we need on password typing times and error rates to be able to accurately measure and reliably predict effects of different password requirements on human performance.

Ultimately, we need data that will help us make informed decisions regarding the institutional implementation costs of various proposed password policies, e.g., if you make change X to the organization's password policy, you can expect approximately a Y increase in the number of password entry errors and associated account lockouts. We must better understand the cognitive and perceptual-motor load on our employees that causes forgotten and/or mistyped passwords before we can confidently estimate the corresponding financial costs on employee time and the IT support staff required to reset locked accounts, etc. To objectively weigh the pros of additional security requirements

against the cons of more onerous passwords for users, we need behavioral data across a series of studies designed specifically to study password rules that mimic those of secure government agencies. To examine effects of changing password policies over time and across devices, we must start by gathering the baseline data necessary for our future comparisons. To this end, we set out to conduct a simple initial study—followed by its replication in a different geographical locale—to collect basic behavioral data examining effects of increasing string length on typing times, errors, and memorability for randomly generated, password-like character strings.

2. PROBLEM BEING SOLVED

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.[1] Using experimental methodology from the behavioral sciences, we initially set out to answer the following questions: 1) How memorable are complex character strings of different lengths that might be used as higher-entropy passwords? 2) How would we differentiate/disambiguate between typing and memory errors? 3) How do we define and measure password memorability? While these were our original research questions, it is important to distinguish them from the methodological problems that we focus on in the current paper. Among those problems: *How do we generate and select stimuli to best examine how manipulating a single element of a multi-faceted password requirement (here, increasing string length) affects memorability?*

3. BACKGROUND AND RELATED WORK

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 [6, 4]. 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 inclusion of upper- and lower-case letters, numerals, and special characters are often recommended for increasing password security [15]. 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.

In order to provide best practice recommendations for institution-wide password policies—regardless of whether the institution is government, academic, or industry-based—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" [14]. It is to be expected that longer passwords should lead to longer entry times and more errors (more characters offer more chances for misremembering and mistyping) but how many characters are too many? When does the burden of remembering become too much for a user and what type of errors do users make when recalling and typing passwords?

There have been many studies of remembering in general (e.g., [16]) and passwords in particular, addressing such issues as memorability, predictability and attention [7, 12, 16, 17, 18, 19, 20]. In addition, there is a large body of literature examining the factors of skilled typing performance from 1923 [5] through the 1980s [9, 11] including a great deal of literature on the cognitive and perceptual-motor aspects of transcription typing [12]. 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., [11]), 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.

4. APPROACH

4.1 Experimental Design

4.1.1 What About Password Policies Are You Studying?

Research comparing and manipulating memorability across passwords is certainly important, but was not the goal of our immediate research. We wanted to focus our initial study on the ramifications of increasing password length requirements only, as length is one of the more commonly used methods of increasing entropy.

4.1.2 Stimuli

4.1.2.1 User-Chosen vs. System-Assigned

Since we could not ask people to use their own passwords due to security and privacy concerns, we had to generate stimuli that would act as stand-ins for "real" passwords. Given our previously described focus on higher-security environments, we chose to use a password generation software program (see Method section for details) with rules configured in alignment with the more stringent of currently enforced and/or newly proposed institutional password policies.

4.1.2.2 Controlling for Stimuli Meaningfulness

Since we only set out to study effects of increasing password length, we wanted to keep other factors constant across stimuli. 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. Rather than making stimuli equally memorable, we wanted them to be equally *unmemorable.* Again, this was intentionally a worst-case scenario and admittedly not representative of how users generate passwords "in the wild."

In addition to controlling for meaningfulness, using randomly generated stimuli should also help address a number of potentially influential variables. For example, some key combinations are physically more difficult to type and some special characters are less familiar than others; some characters are more visually

[1] This is not intended to be an exhaustive discussion of all facets of password memorability. We focus here on those aspects most relevant for the current paper; tasks such as password creation were not included in this work and we do not address them here.

similar than others. Rather than explicitly manipulate and control for such factors, it was considered that randomization would account for variables like these.

4.1.2.3 Stimuli Order

All participants received the same strings in the same random order. Had we used a different set of stimuli for each participant, we would not have been able to separate subject effects from password effects. Had we used a different random order per participant, it would have necessitated additional programming complexity on the front-end implementation, as well as extra complexity of back-end data parsing. This additional level of complexity did not seem warranted given the simple nature of these first studies; we assumed that a randomly determined order would be random enough. This assumption may not be true and we may actually want to have a different random sequence per participant.

4.1.3 Stimuli Memorization

4.1.3.1 Practice

Once we generated our stimuli and decided on presentation order, we had to decide how we would allow participants to practice. We chose practice-at-will rather than enforced practice (e.g., forcing a specific number of error-free repetitions). Because we did not use enforced practice, it was necessary to provide a verification step to help ensure the participant had memorized the string.

4.1.3.2 Recall

While verification was used to test memorization of individual strings, a surprise recall test was used to test recall across all strings at the end of the study.

4.2 Programming

Because timing was one of the main dependent variables, we had to be sure that the method we used to collect timing didn't unduly affect the recorded times. Therefore, we compared hooking versus application-level logging. The difference was much smaller than what the fastest typist could type. As application-logging also gave us what field participants were in when they pressed a given key, we chose that method. In addition to capturing timing and entry field, we also needed to record whether participants were in the practice or test phase, and whether the entered string matched the target string.

4.3 Pilot-testing

We used the following eight strings—two each of lengths 6, 8, 10, and 12—during pilot testing with four participants:

```
w0vM5i
s58xTBU,d1
i4M1a4Po
k=80yDO&U2Hf
c7MV6w
u54FbKMe*8
r5]1xWH0NVs>
m2n1E4Xo
```

All participants received the same eight strings in the same random order shown above. We found that simply using password generation software (described in detail in the Method section) was insufficient to ensure equally meaningless stimuli. Of the eight strings used during pilot testing, one stood out in particular: three of the four pilot participants commented that the center part of the `u54FbKMe*8` string looked like "Facebook Me." The difference in meaningfulness could account for the unexpected increase in accuracy observed for three of four participants at string length 10. Not only did this string have more inherent meaning than the other pilot strings, but what should have been the more challenging components, i.e., the numbers and special symbol, were fairly easy to type: five and four are on adjacent keys, and the asterisk and the number eight are on the same key. In addition to emphasizing the need to control for differences in meaningfulness across stimuli, this one string illustrates another significant methodological challenge: the difficulty of disentangling memory errors from typing errors in single-session laboratory studies where it is unlikely that participants can practice any or all strings sufficiently to transition them from declarative to procedural memory.

5. METHOD

5.1 Participants

Two studies were conducted. The first consisted of 31 participants recruited from the Washington DC (WDC) metropolitan area. The second study replicated the first with a new group of 45 participants recruited from the University College London (UCL). There were 76 participants in total. Participant ages in the WDC study ranged from 18 to 78 years, while participant ages in the UCL study ranged from 18 to 27 years.

5.2 Design

A within-subjects design was used, where all participants typed the same ten strings (two strings each of six, eight, ten, twelve, and fourteen character lengths) ten times each.

5.3 Instructional Materials

Participants were issued the following verbal instructions at the beginning of the experiment:

"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. 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."

5.4 Apparatus and Instrumentation

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. PCs in each study ran the same custom software program—designed in-house for the experiment—that presented character strings to participants one at a time. The program allowed participants to enter the strings as described in the following Procedure Section, and created a separate log file for each participant. For both downstrokes and upstrokes, a log entry containing a timestamp (in ms) and the key's identity was recorded.

5.4.1 Stimuli

The strings participants were asked to memorize consisted of ten strings total: two strings each of six, eight, ten, twelve, and

fourteen character lengths (the additional 14-character length was suggested by our statistician based on pilot-testing results). Strings were randomly generated using commercially available password generation software[2]. Strings had to consist of at least one upper-case character, one lower case character, one number, and one special character; they could not begin with an upper-case character, nor could they end with an exclamation mark. The ten strings were as follows:

```
5c2'Qe
m#o)fp^2aRf207
m3)61fHw
d51)u4;X3wrf
p4d46*3TxY
q80<U/C2mv
6n04%Ei'Hm3V
4i_55fQ$2Mnh30
3.bH1o
ua7t?C2#
```

Stimuli were presented in the font shown above, Consolas, 18 point, whereas any other text in the data collection program was presented in Calibri. The text participants typed during the practice, verification, and recall phases (see Procedure section) of the experiment was also shown in Consolas. We chose the Consolas font for the target strings and participant-typed text because it clearly distinguishes between the letter "o" and the number zero: note the diagonal line through the zeros in the stimuli above. Monospace fonts like Consolas are frequently used in programming in order to disambiguate commonly confused characters (e.g., the number zero vs. lowercase o, and lowercase l vs. uppercase I vs. the number 1). The order of the strings shown above is random. They were presented in the sequence shown above for all participants in the first (WDC) study. The order of the last two strings was reversed for the second (UCL) study due to a software configuration file change.

5.5 Procedure

Each participant was given the verbal instructions described in the Instructional Materials section, after which they were given an informed consent form to sign. The test facilitator then started the data collection program and entered the participant number. From that point, the participant had one hour to complete the task described in the instructions (as well as a surprise recall task described at the end of this section). For each of the 10 character strings, participants saw a series of three screens: practice, verification, and entry screens.

Instructions on the practice screen asked participants to memorize the target string (Figure. 1). Participants were free to practice the string as many or as few times as they wanted.

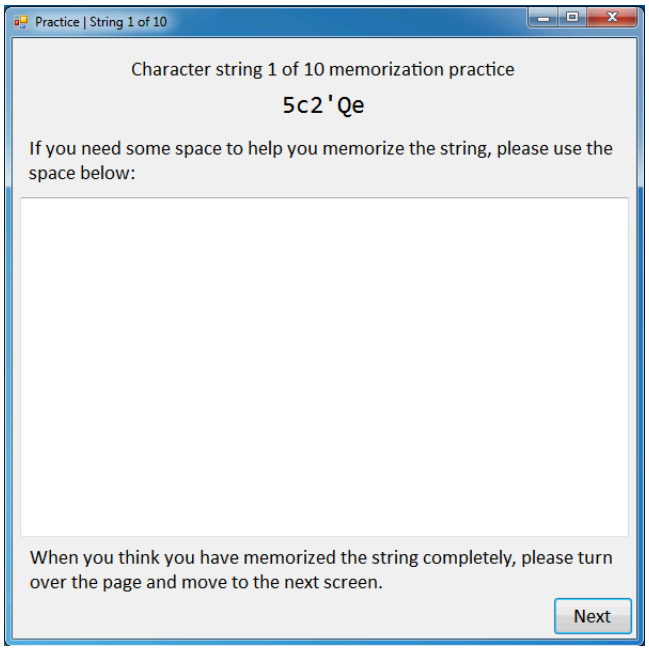

Figure 1. Practice Screen.

When participants felt they had sufficiently practiced and memorized the target string, they moved on to the next screen for verification (**Figure 2**).

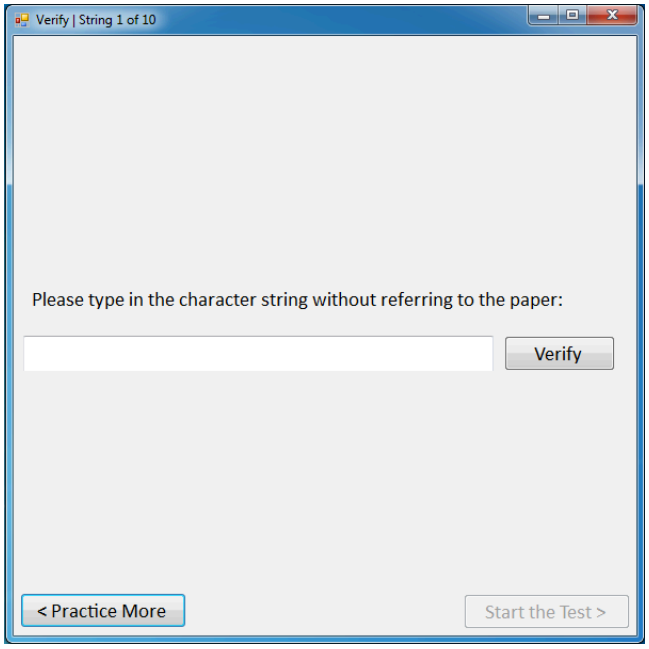

Figure 2. Verification Screen.

The verification screen asked participants to type the memorized target string. If the typed string failed the verification – in other words, if it did not *exactly match* the string the participant had been asked to memorize – the participant could either go back to the practice screen or try and verify the string again. On the

[2] Advanced Password Generator from BinaryMark was used, http://www.binarymark.com/Products/PasswordGenerator/default.aspx 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.

verification screen, participants had to enter the string correctly in order to move to the third screen, where they were asked to enter the memorized string ten times (**Figure 3**).

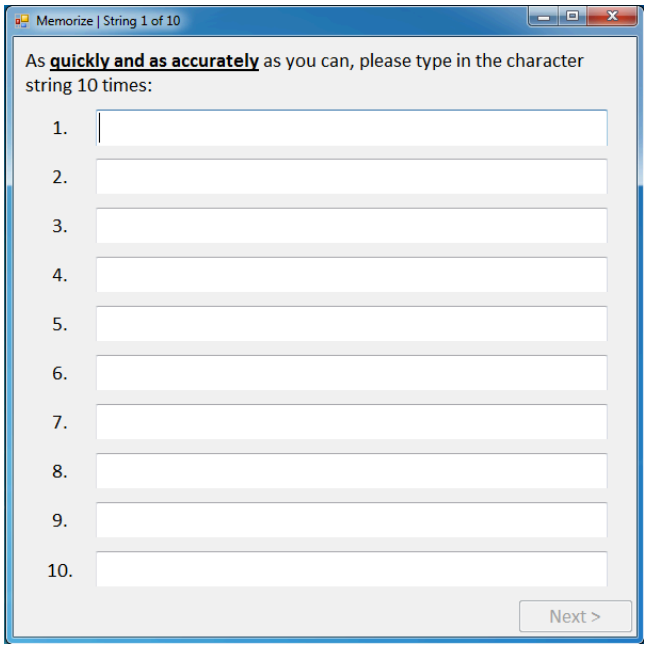

Figure 3. Entry Screen.

This procedure (practice at will, verify correctly once, enter string 10 times) was repeated for each of the ten strings. After all ten strings had been tested, the program gave the participants a surprise recall test to see how many of the ten strings they remembered. Instructions on the surprise recall screen read "Please type in as many of the character strings as you can. (Note: they don't have to be in the order you received them)." Aside from the instructions, the recall and entry screens were nearly identical.

Typed text was visible during practice and verification (Figures 1 and 2, respectively), masked with asterisks during entry (Figure 3), then visible during surprise recall (not shown given its high similarity to the entry screen, Figure 3).

6. DATA COLLECTION

The program recorded the time participants took to practice, verify, and enter each individual string and whether or not the participants' entered string matched the target string. Researchers reviewed the data for the number and type of errors participants made while entering each individual string, as well as the number of strings each participant recalled correctly during the surprise recall task.

7. ANALYSIS

There are a number of ways one could analyze and present data like these. Determining which timings are theoretically most meaningful depends on the question you are trying to answer. One can examine only error-free string entries vs. all entries; report string entry times with their associated cognitive processing time, etc. Given our definition of password memorability (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), here we examined times across those experimental phases.

8. RESULTS

In this section, we present the number of entry errors per string length (lengths of 6, 8, 10, 12, and 14 characters); the average time taken by participants across practice, verification, and entry phases (again, per string length); the types of errors made during string entry; and the number of strings participants recalled correctly during the final surprise recall task. For purposes of the experiment, an entered string was considered to be an error if it did not exactly match the target string upon submittal. Unless otherwise noted, timings were collapsed across each string length category

8.1 Errors for Individual Entry Tasks

Figure 4 and Figure 5 show the median errors per character string for the WDC and UCL studies, respectively. UCL participants made fewer median errors overall than did WDC participants for strings ten or more characters in length. In both studies, participants had increases in the variability of errors as the string length increased.

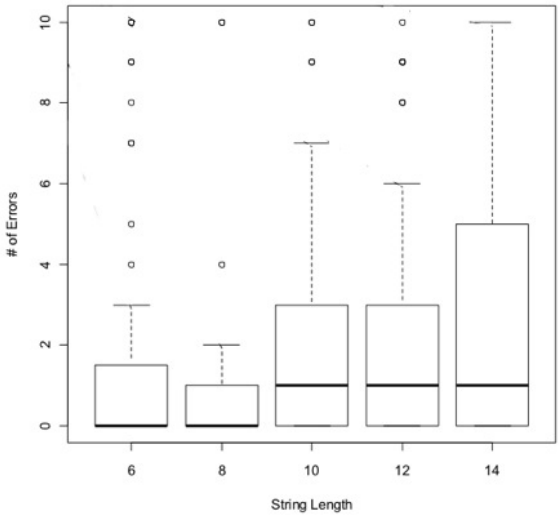

Figure 4. Median Number of WDC Errors by Target String Length.

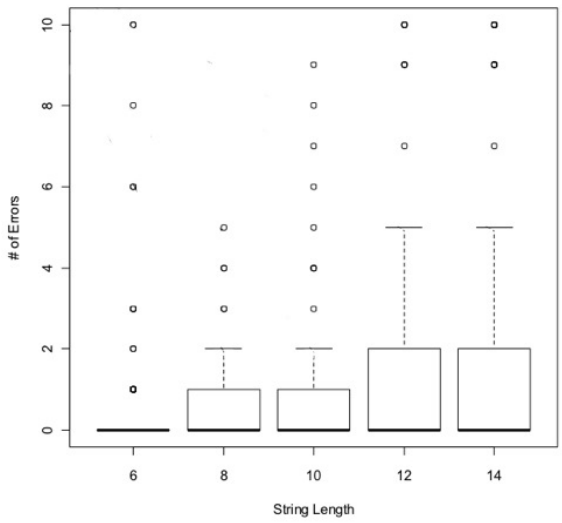

Figure 5. Median Number of UCL Errors by Target String Length.

8.2 Times Across Practice, Verification, and Entry Tasks

Seven of the 31 participants in the WDC study failed to complete the test in the one-hour time allotted. In the UCL study, all 45 participants completed the test within the one-hour time limit. The time each participant took per string across practice, verification, and entry phases was calculated from the point when the practice screen was first presented (**Figure 1**) to when the tenth repetition of said string was typed on the entry screen (**Figure 3**). These times include both correct and incorrect entries, and provide a global view of the general memorability of the various strings (i.e., difficulty across the practice and entry phases).

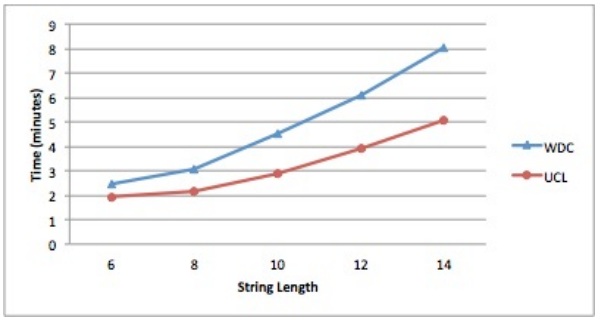

Figure 6. Mean Per String Length Times Across Practice, Verification, and Entry Phases.

As **Figure 6** shows, the average time taken to practice, verify, and repeatedly enter a string was monotonically increasing as the length of the character string increased. The figure also shows that the rate of increase was greater for strings eight characters or more in length for both studies (with the WDC study participants taking somewhat longer overall).

8.3 Times for Entry-Only Tasks

Whereas times presented in **Section 4.2** span multiple test phases (practice, verification, entry), times in this section are presented for the entry phase only. To compute the mean entry time for a particular character string, the total time spent on that string's entry screen was divided by 10 (the number of string entry fields, as shown previously in Figure 3). These entry phase times include cognitive processing time (to retrieve a string and initiate typing); typing times (including any error corrections made by participants); and transition times between the 10 entry fields (participants could navigate between text fields either by using the mouse or tabbing).

Tables 1 and **2** present string entry times for the WDC and UCL studies respectively; as in previous sections, times are collapsed for each of the string length categories. In contrast, **Tables 3** and **4** (again, for WDC and UCL respectively) present entry times for each of the ten strings individually; this lower level of granularity better illustrates timing differences between specific strings. As previously mentioned, seven WDC study participants did not complete the study in its entirety, only making it through the fifth string; their entry times are not included for strings six through ten.

Table 1. WDC entry times (seconds) by string length.

String Category	Min	Max	Mean	SD
Length 6	36	105	65	21
Length 8	44	130	83	26
Length 10	54	187	113	36
Length 12	69	234	144	42
Length 14	81	329	170	61

Table 2. UCL entry times (seconds) by string length.

String Category	Min	Max	Mean	SD
Length 6	34	98	61	15
Length 8	30	99	55	14
Length 10	51	211	88	31
Length 12	57	211	113	32
Length 14	68	280	135	44

Table 3. WDC individual string entry times (seconds).

String	Min	Max	Mean	SD
5c2'Qe	38	138	80	27
m#o)fp^2aRf207	86	663	208	114
m3)61fHw	39	144	83	24
d51)u4;X3wrf	68	238	141	42
p4d46*3TxY	51	250	122	45
q80<U/C2mv	56	217	116	44
6n04%Ei'Hm3V	70	326	157	61
4i_55fQ$2Mnh30	71	303	158	58
3.bH1o	30	90	56	18
ua7t?C2#	49	142	87	30

Table 4. UCL individual string entry times (seconds).

String	Min	Max	Mean	SD
5c2'Qe	27	114	56	18
m#o)fp^2aRf207	73	308	138	50
m3)61fHw	34	130	66	18
d51)u4;X3wrf	50	174	106	29
p4d46*3TxY	42	174	82	25
q80<U/C2mv	49	248	94	40
6n04%Ei'Hm3V	63	248	119	39
4i_55fQ$2Mnh30	63	311	132	48
3.bH1o	35	114	67	18
ua7t?C2#	25	81	43	14

8.4 Types of Entry Errors

Each erroneously entered string was analyzed regarding the nature of the error(s) it contained. The types of errors made were as follows:

- Extra character
- Wrong character
- Incorrect shifting (i.e., entering a shifted character when not required or a non-shifted character when a shifted character was required)
- Transposition of characters (i.e., switching the order of two adjacent characters in the target string)
- Character was in the wrong place within the string (this does not include transposition of characters as described in the bullet point above)
- Character typed was adjacent on the keyboard to the target character (e.g., "s" instead of "a")
- Zero instead of an "O"
- Missing character

The error of typing a zero rather than an "O" was so common that we decided to make this error a separate item from the more generic "wrong character" error. **Table 5** shows the percentage of each type of error in the WDC participant group, UCL participant group, and combined participant group. In total, the WDC group made 471 errors and the UCL group made 556, for a grand total of 1,027 errors.

Table 5. Error types and percentages during string entry.

Error Type	WDC	UCL	All
Extra character	7%	7%	7%
Wrong character	6%	12%	9%
Incorrect shifting	38%	51%	45%
Transposition of characters	10%	6%	8%
Adjacent keyboard character	8%	10%	9%
Zero instead of "O"	3%	3%	3%
Missing character	25%	10%	17%
Character misplaced within string	3%	1%	2%

As seen in Table 5, the most common type of error was incorrect shifting (45%), followed by missing characters (17%).

8.5 Surprise Recall Task

A string was considered recalled correctly only if it exactly matched a target string. In the surprise string recall task, nine participants could not remember any strings, while most could remember only one or two strings (32 remembered one and 22 remembered two, where 32 and 22 are mutually exclusive counts). The largest number of strings recalled correctly by any participant was four – and only one participant managed to do so (see **Table 6**).

Table 6. Number of strings remembered by participants during surprise recall task (numbers represent counts).

Strings	WDC	UCL	All
0	6	3	9
1	11	21	32
2	5	17	22
3	2	3	5
4	0	1	1

The most frequently recalled string during the surprise recall task was the last string memorized, followed by the second-to-last string memorized for each study (see Table 7).

Table 7. Most frequently remembered strings during surprise recall task (numbers represent counts).

String	WDC	UCL	All
ua7t?C2#	17	21	38
3.bH1o	7	33	40
q80<U/C2mv	2	2	4
4i_55fQ$2Mnh30	1	10	11
6n04%Ei'Hm3V	1	1	2
m#o)fp^2aRf207	0	1	1

9. 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 7**). 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 studies, even though UCL participants were faster overall (see

Figure 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 study 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 study data (see **Figure 4**), where the median number of errors increased from zero to one between the eight- and ten-character strings. The variability of the errors increased at the same point. Even though the median number of errors remained the same for the UCL study participants (see **Figure 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 studies, the difference in the "error variability threshold" may potentially be explained by the younger UCL study participants having greater memory capacity and/or better general typing skills. To test these potential explanations, future studies could 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 studies 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 studies, the largest percentage of errors was shifting errors (see **Table 5**). Since many special characters require a shift action (e.g., "8" must be shifted to "*"), these errors are particularly important given the increasing use of special characters in password policies.

10. LIMITATIONS

10.1 Internal Control vs. External Validity

There are always tradeoffs between tight experimental control and generalizability; this tension is certainly not unique to the usable security community and clearly exists across other research domains as well. It is possible to exercise greater control for potentially confounding factors in laboratory studies than it is "in the field," yet their higher levels of internal control typically come with associated costs to external validity and generalizability. Such was the case with the currently reported laboratory studies, where we chose to conduct basic research examining a commonly threatened worst-case scenario (at least from the point of view of the user), in which people are forced to use complex system-generated passwords. While our stimuli were representative of system-generated passwords in higher-security settings, there were other aspects of our studies that were less realistic. Even in higher-security environments, users are often allowed to choose their password from a short list of system-generated passwords; they are encouraged to pick the one they can most easily remember. Regardless of whether it is a maximum or minimum security environment, it is unlikely that users in the "real world" would have to learn 10 new passwords in a single one-hour session like ours. It is more often the case that account passwords expire at uneven intervals [4].

10.2 Randomly Generated Stimuli

We found that using randomly generated stimuli wasn't as straightforward as originally anticipated; while using complex password generation software would theoretically allow the study to enjoy the benefits that typically accompany random stimuli generation, in reality, the number of stimuli was simply too small to exploit the benefits of randomization. While we could have increased the number of strings used, this would have necessitated lengthening an already difficult testing session or bringing users in for multiple sessions.

While we set out to investigate effects of increasing string length only, we found that we could not fully disentangle these from other string components. In these studies, the frequency and difficulty (i.e., requiring a shift or not) of special symbols was unevenly distributed across the different string lengths we studied; purely by chance, shorter strings contained easier symbols. Due to this confound, we were unable to fully disambiguate effects of string length relative to those of special symbols. While it would certainly be feasible to better control for the effect of special symbols relative to length, we argue that it would be nearly impossible to address all other variables simultaneously, such as the rarity of certain special characters relative to others. While people use periods, exclamation marks, and apostrophes regularly in typing normal prose, other symbols such as a percent sign or equal sign are less common. We did not control for or objectively measure participants' familiarity with some symbols relative to others; this could easily be addressed by having some type of number- and symbol-specific baseline typing task to gauge participants' typing skills for those more problematic characters.

We were also surprised by the degree to which the attempt to control for password meaning—by making all stimuli equally meaningless via random generation—failed. This was clear merely based on four pilot participants; even after replacing the problematic stimuli found during piloting, it became clear based on study participants' comments that the new stimuli were not all created equal. For example, multiple participants mentioned remembering the string containing "3V" by associating it with "three-volt". It was simply easier for subjects to assign meaning to some strings more so than others. This issue is not unique to our study; even studies that intentionally manipulate memorability can't possibly control for the fact that some strings are more meaningful to some individuals than others. For example, participants born in 1980 may find remembering the string beginning with "q80" easier than those born in 1983.

10.3 Data Collection

Had we collected additional demographic information, we could have expanded the types of questions we could explore with our data. Additional data that could have been collected include:

- Age (exact rather than ranges)
- Typing ability
- Education
- Cognitive ability
- Subjective rating of difficulty
- Strategy

Since our focus was on password policy—and institution-wide password policies do not differ based on individual characteristics like age—we did not collect certain demographic data. While those data would not affect the conclusions of our work from a

practical policy perspective, upon retrospect, we wish we would have collected them from a purely research perspective.

10.4 Analysis

10.4.1 Complexity of Data Files

Another challenge was the complexity of the raw data files. The variability in participant strategies (differing amount and type of practice, varying number of verification attempts, etc.) caused the data files to become quite complex. This complexity necessitated additional programming simply to reformat the data in the manner needed for analysis. One example of the difficulty caused by variability in participant strategies was the issue of carriage returns.

10.4.1.1 Carriage Returns

One unforeseen consequence of recording keystrokes came when we transferred the keystroke-timing file to a spreadsheet application. We expected that each keystroke would appear on a separate row with its timing, name, character, and whether it matched the target string data, and this was true for most of the characters typed. The surprise came when a participant typed an "Enter" keystroke. This keystroke was recorded, properly, as a carriage return. When imported into the spreadsheet application the carriage return caused the spreadsheet to deposit the data after the carriage return, on a new row. Going to a new row, within one keystroke line of data, was especially problematic when the participant typed "Enter" many times, causing one keystroke's data to be spread over many rows. To address the unintentional row addition, a software program was implemented to replace all carriage returns with a blank. The data file would still include the keystroke name, but not the actual keystroke.

11. CONCLUSIONS AND FUTURE WORK

Since capturing real-world password typing data poses significant privacy and security concerns, we instead gathered human performance data in two controlled laboratory experiments 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 both our results and our methodology.

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. Will methodological decisions made in the desktop research environment port well to the mobile platform?

The current research is a first step 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 passwords may we hope to predict how well users will be able to comply with proposed password policy changes. These two studies contribute to the usable security community by presenting much-needed (albeit, unsurprising) human performance data that are difficult to obtain in the real world. More importantly, they illustrate several methodological challenges that other researchers in this space may benefit from considering more carefully in future studies.

We learned that a simple research question does not necessarily mean a simple study; it is important to keep methodological rigor high for even the simplest of study design decisions. A seemingly small decision or assumption can have large ramifications on subsequent conclusions. We assumed that randomly generated stimuli should control for attributes like meaningfulness, and also ensure that special symbols were fairly evenly distributed across strings. While this may be true over a large number of strings, purely by chance, the subset chosen for our study unfortunately had special characters that were not so special, i.e., a period and an apostrophe for the two strings of length six. This meant we could not fully determine effects of increasing string length in isolation from increasing complexity of special symbols. We could not disassociate memory from typing errors with the current data set. Since it is unlikely that participants in single- or even multi-session laboratory studies ever learn experimental stimuli as well as they know their own passwords, disambiguating typing from memory errors in laboratory studies remains challenging. It is possible that alternative methods, such as computational cognitive modeling [1], can augment our behavioral studies. With a well-validated model at the level of individual users, we may eventually hope to scale our modeling up to the larger institution-wide predictions described in this paper's introduction section.

Together, these findings and their accompanying methodological discussion contribute significantly to the usable security research community by helping address a much larger question: how do we best study passwords in laboratory experiments where we cannot replicate the real-world environment in which users choose their own passwords?

12. ACKNOWLEDGEMENTS

The authors gratefully acknowledge Ross J. Micheals for instrumenting the custom data collection software; Hung-kung Liu for statistical consultation and analysis; and Terry Brugger and Paul Young-Jin Lee for their help in post-study data processing. This work was funded by the Comprehensive National Cybersecurity Initiative (CNCI).

13. REFERENCES

[1] Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111* (4), 1036-60.

[2] Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. Bower (Ed.), *Recent advances in learning and motivation (Vol. 8)* (pp. 47-90). New York, NY: Academic Press.

[3] Chiasson, S., Forget, A., Stobert, E., Van Oorschot, P., & Biddle, R. (2009). 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).

[4] Choong, Y., Theofanos, M., & Liu, H. (2013). Federal employees' password management behaviors – a Department of Commerce case study. (Manuscript in preparation).

[5] Coover, J. E. (1923). A method of teaching typewriting based upon a psychological analysis of expert typing. *National Education Association*, 61, 561-567.

[6] Florencio, D., & Herley, C. "A large-scale study of web password habits," in WWW 2007, (Banff, Canada, 2007), ACM Press.

[7] Forget, A., & Biddle, R. (2008). Memorability of persuasive passwords. *CHI'08 Extended Abstracts on Human Factors in Computing Systems* (pp. 3759–3764).

[8] Gehringer, E. F. (2002). Choosing passwords: Security and human factors. *International Symposium on Technology and Society, 2002.*(ISTAS'02). (pp. 369–373).

[9] Gentner, D. (1981). Skilled finger movements in typing. Center for Information Processing, University of California, San Diego. CHIP Report 104.

[10] Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81-97. doi: 10.1037/h0043158

[11] Salthouse, T. (1984). Effects of age and skill in typing. *Journal of Experimental Psychology, Vol. 113*, No. 3, 345-371.

[12] Salthouse, T. (1986). Perceptual, cognitive, and motoric aspects of transcription typing. *Psychological Bulletin, Vol. 99*, No. 3, 303-319.

[13] Shay, R., Kelley, P. G., Komanduri, S., Mazurek, M. L., Ur, B., Vidas, T., Bauer, L., Christin, N., & Cranor, L. F. (2012). Correct horse battery staple: Exploring the usability of system-assigned passphrases. *Symposium on Usable Privacy and Security (SOUPS) 2012,* Washington, DC.

[14] United States Department of Commerce, National Institute of Standards and Technology (NIST). (1985). *Password usage* (FIPS PUB 112). Retrieved from website: http://www.itl.nist.gov/fipspubs/fip112.htm

[15] United States Department of Homeland Security, United States Computer Emergency Readiness Team (US-CERT) (2009). *Security tip (ST04-002): Choosing and protecting passwords*. Retrieved from website: http://www.us-cert.gov/cas/tips/ST04-002.html

[16] Unsworth, N., & Engle, R. W. (2007). The foundations of remembering: Essays in honor of Henry L. Roedgier III, 241–258. New York: Psychology Press.

[17] Vu, K., Bhargav-Spantzel, A., & Proctor, R. (2003). Imposing password restrictions for multiple accounts: Impact on generation and recall of passwords. HFES 47th Annual Meeting (pp. 1331-1335).

[18] Vu, K., Cook, J., Bhargav-Spantzel, A., & Proctor, R. W. (2006). Short- and long-term retention of passwords generated by first-letter and entire-word mnemonic methods. *Proceedings of the 5th Annual Security Conference*, Las Vegas, NV.

[19] Vu, K., Proctor, R., Bhargav-Spantzel, A., Tai, B., Cook., J., & Schultz, E. (2006). Improving password security and memorability to protect personal and organizational information. *International Journal of Human-Computer Studies, 65*, 744-757.

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