BREAKING: Password Entry Is Fine

Catlin Pidel(✉) and Stephan Neuhaus

Zurich University of Applied Sciences, Zurich, Switzerland
{catlin.pidel,stephan.neuhaus}@zhaw.ch

Abstract. In our digital world, we have become well acquainted with the login form—username shown in plaintext, password shown in asterisks or dots. This design dates back to the early days of terminal computing, and despite huge changes in nearly every other area, the humble login form remains largely untouched. When coupled with the ubiquity of smartphones, this means we often find ourselves entering complex passwords on a tiny touchscreen keyboard with little or no visual feedback on what is being typed. This paper explores how password masking on mobile devices affects the error rate for password entry. We created an app where users entered selected passwords into masked and unmasked password fields, measuring various metrics such as typing speed, error rate, and number of backspaces. We then did an exploratory analysis of the data. Our findings show that, perhaps unexpectedly, there is no significant difference between masked and unmasked passwords for any of these metrics.

Keywords: Security · Passwords · Data entry errors · Mobile security · Mobile usability

1 Introduction

The average person logs into 7–25 accounts every day [4], and the vast majority of people memorize their passwords, meaning reuse of identical or similar passwords is common [13]. This stands in opposition to traditional security advice, which dictates that passwords should be long, random strings full of complex characters, changed regularly, and never reused across accounts. This can really only be done by using a password manager, which only 3% of Americans use consistently [13]. What is more, in the smartphone era, tiny keyboards with separate screens for letters and special characters can turn a strong password into a usability nightmare: mobile users mistype their passwords twice as often as desktop users, and mobile password entry also takes 20% longer [12].

Password unmasking is the practice of displaying passwords rather than masking them with asterisks or dots. Is this a viable alternative? Usability experts claim that masking passwords is an unnecessary complication that causes confusion and frustration. But some security experts warn against the dangers of shoulder surfing (stealing passwords by looking over people's shoulders as

A. Moallem (Ed.): HCII 2019, LNCS 11594, pp. 67–80, 2019.
https://doi.org/10.1007/978-3-030-22351-9_5

they type). While there are a number of articles and blog posts espousing one view or the other, there is little evidence on how much password masking actually hinders usability. This paper aims to take on a small part of this overlying issue: how does the error rate for password entry compare between masked and unmasked passwords, specifically for smartphone users?

1.1 The Origins of Password Masking

Password masking comes from terminal computing, when every command was printed out on paper [9]. In those days, it made sense not to echo passwords at all, or at least to replace them with asterisks. Without it, anyone with access to the computer printouts could easily harvest someone else's login info. However, the days of such paper trails are long gone, and this rationale no longer applies. This begs the question: does password masking still serve a purpose, or is it simply there because it always has been?

1.2 Contributions

Despite having been pronounced dead many times [9], passwords are the most used authentication mechanism on the Internet today [10]. Similarly, mobile phones are ubiquitous, so gaining a deeper understanding of entering a password on a mobile device is a relevant issue. For as long as passwords have been around, we are still in a "data poor" research state [1,10], so this study contributes to password research as well as the growing intersection between usability and security research. We specifically make the following contributions:

- We compare the error rates of different types of passwords on mobile phones.
- We tested two different types of password fields: masked and unmasked.
- Additionally, we tested these logins with different types of passwords to see what role the password length and density of special characters play in this.

2 Related Work

2.1 Usability and Security of Password Masking

In the early days of computers, password masking helped communicate that a password was sensitive information [19], but this is now common knowledge, and not necessarily an argument for its continued use. From a modern usability perspective, the lack of visual feedback makes it difficult to find and correct mistakes in masked passwords, leading to unnecessary frustration. Usability expert Jakob Nielsen [15] and security expert Schneier [19] have argued that masking passwords causes users to choose easier (and therefore less secure) passwords, reuse passwords across accounts, or even copy and paste them from a file.

That said, password masking is also what the user has come to expect. A small-scale user study showed that unmasking passwords undermined the user's trust—even though masking is a purely cosmetic fix, 60% of participants became

suspicious of a site when their passwords were displayed in plaintext [11]. There's also a perceived security threat of unmasking passwords. What if someone looks over your shoulder without your knowing? Additionally, how do you navigate the trust dynamics of entering a password with someone else present? Is it alright to ask your boss, your partner, or your child to avert their eyes from your login form [19]?

The most common security threat associated with plaintext passwords is shoulder surfing. Both Schneier and Nielsen have downplayed this risk, a point which many of the commenters on these articles disagreed with [15,19]. It is much harder to snoop a password typed onto a smartphone screen, and with mobile internet usage recently overtaking desktop traffic [4], the threat of shoulder surfing is becoming less common compared to the nuisance of repeatedly typing a password [20]. However, dissenters argue that there are still plenty of scenarios where password masking provides a necessary layer of security. Using a computer in a public setting like a coffee shop, for example, as well as needing to enter a password during a presentation, are legitimate use cases where plaintext passwords fall short.

Despite the considerable debate, we were unable to find research directly comparing the error rate between masked and unmasked password fields.

2.2 Usability and Security of Different Types of Logins

There is a fair amount of research into alternatives for the username and password combination, such as the usability of click-based graphical passwords [2] and password managers [3]. One study generated a "profile" of how the user types their password, based off of the time a key is held down and the time between key presses [17]. When the profiles of valid users were compared against those of impostors entering the same password, this additional metric helped filter out unauthorized users. However, typos and the subsequent use of backspace interfere with this metric, making it infeasible for large-scale adoption. Another study explored an alternative to password masking, the TransparentMask [8]. This combines the typical black dots with symbols that represent a hash of the last n characters of a password. The idea is that since humans can easily recognize sequences of symbols, it would provide a way to alert the user to typos without providing as much of a security risk as a fully unmasked password (Fig. 1).

While these ideas are fascinating, the likelihood for their widespread adoption is unlikely, and may also lead to user confusion. According to one survey of web tools [14], login forms are among the least standardized website components. Introducing new paradigms, however well-intentioned, may only serve to muddy the waters. Additionally, while the typical login form is not as interesting, its persistence in our digital world demands more research than has currently been done on the subject [1,10].

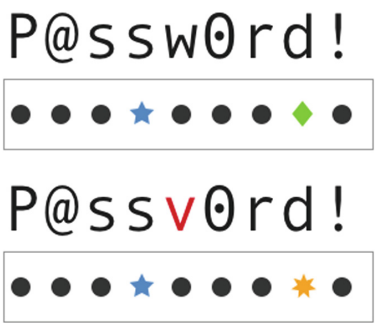

Fig. 1. Transparent Mask example inspired by Gruschka and Iacono's paper [8]. The colorful symbols (blue star, green diamond) represent a hash of the previous characters in a password. A typo in the preceding characters would result in a different symbol appearing, as shown when the "w" is mistyped as a "v". (Color figure online)

2.3 Mobile Password Entry

As mobile phone usage continues to increase [4], passwords are increasingly being entered on mobile devices. And while mobile devices are often cited as a reason why password unmasking is important (tiny keyboards leading to an increased likelihood of typos, for example), we did not find studies measuring how much masked password fields affect usability on mobile devices. One dissertation provides an in-depth analysis of password entry on mobile devices—what influences the user typing speed (such as switching keyboards to find special characters), and the most commonly made mistakes [5]. Prior research also explored the error rate of typing passwords on mobile phones compared to desktops [12], as well as an analysis of the "number and nature of errors committed during password entry" and whether the user notices these errors before submitting the password [7]. Our work provides another facet to the question of mobile password entry by studying the error rates for both masked and unmasked login prompts.

From a security perspective, Schaub et al. [18] examined how the design of different smartphone keyboards (iOS, Windows, Android, and others) can make shoulder surfing easier or more difficult. There has also been research into how the platform (mobile, tablet, desktop) affects the makeup of the password created [21], but not whether certain types of passwords are more or less error-prone in daily mobile use.

3 Methodology

This study aims to answer two questions: how password masking affects the error rate of password entry on mobile devices, and how the makeup of a password influences the aforementioned error rate.

3.1 Effects of Password Masking

We tested two different login forms—one with typical mobile password masking (where the password is masked except for the most recently typed character), and a fully unmasked password, displayed in plaintext. We also considered testing a third option where the user has the choice to mask or unmask their password with a checkbox, but we ultimately decided that this was better tested in a separate study concerned with whether users will choose to change the default masking of a password.

3.2 Effects of Password Makeup

There is a lot of variation in passwords, but we have identified four main categories:

- **Pass phrases**: multiple words separated by spaces ("`cats are fantastic friends`")
- **Typical passwords**: a single (in our case, English) word with letters replaced by numbers and special characters ("`C@terp!11ar2018`").
- **Randomized**: a fully randomized string of characters, such as those used by password managers ("`nBqzEcP2A}Q8,jG`")
- **Bad passwords**: passwords often seen in password leaks ("`password123`")

Of these four types, we are not interested in "bad passwords" because they have already been shown to be a security threat. For the 3 more secure alternatives, we generated a list of 10 passwords at equivalent password strength (as measured by our chosen software, 1Password). Our hypothesis is that the error rate for password entry varies both by type of password and by whether the password field is masked.

3.3 Study Structure

Our study measures the error rate of password entry using a custom-built app, PasswordResearcher[1]. This app is built in Unity and deployed to both iPhone and Android, and is available with either English or German keyboards depending on the preferences of the user. This is to mitigate any errors that could arise from using an unfamiliar keyboard layout.

Each session has two parts: one half where the passwords are masked and the other with passwords unmasked. We alternate whether the participant starts with the masked or unmasked task each time. To ensure there is no overlap of passwords between these two portions, the app randomly chooses four passwords from each of the three categories and splits them into two groups of six passwords (two from each category). For each half of the study, the participant sees each given password three times with all passwords shuffled in a random order. Entering the password multiple times is intended to elicit a learning response

[1] https://github.com/catiejo/PasswordResearcher.

since participants learn to type their passwords more quickly and consistently over time. Additionally, the random shuffling allows us to test all three password types under similar conditions—it controls for participants getting bored and therefore less careful over time, or alternatively, getting more used to typing and getting better.

The passwords are displayed on the screen as an image so it is not possible to copy-paste the on-screen password, and also so the user does not need to memorize the password. We were careful not to mention that we were interested in the error rate of password entry since this could change their natural behavior and skew our results.

3.4 Data Collected

We used a recruiting agency, TestingTime[2], to recruit ten participants. For each password attempt, we recorded:

- Expected password
- Actual (typed) password
- Entry time
- Password type (phrase, random, or typical)
- Participant ID
- Overall attempt number (1 to 18 for each half)
- Password attempt number (1 to 3)
- Operating system (Android or iOS)

We also collected qualitative exit questions from each participant:

- Do you remember any of the passwords you typed?
- Which of the password types, if any, did you feel were quicker to type?
- Which of the password types, if any, did you feel were easier or more difficult to type?

It is worth noting that this study is only concerned with whether a login attempt would be successful or not, so we are measuring errors on submission rather than incremental errors made while typing. The latter has potential for a follow-up study, but is outside the scope of the present one.

4 Results and Analysis

We defined an incorrect password attempt as one where the actual password entered by the user does not match the expected password. We used the Levenshtein edit distance to further quantify the correctness of a password and to give a lower bound on the actual number of mistakes that were made by the user.

After collecting the data, we removed any attempts that took more than 100 s. This happened occasionally when participants set down the phone to come ask

[2] https://www.testingtime.com/en/.

a question. Similarly, we removed any attempts with more than ten mistakes, which sometimes happened when participants accidentally pressed the enter key too soon.

Using this data, we calculated the error rates, which we defined as the total number of incorrect attempts (where the expected and actual passwords are not the same) divided by the total number of attempts for each of the following metrics:

- Masked error rate: 18%
- Unmasked error rate: 15%
- Phrase password error rate: 15%
- Typical password error rate: 22%
- Random password error rate: 12%
- Overall error rate: 16%

While there are differences in these error rates (especially between random and typical passwords), the pseudoreplication and small sample size of this study prevent them from being statistically significant. We therefore opted to explore and visualize results without trying to prove statistical significance (see Sect. 6). The conclusions below are our subjective interpretation of the data.

4.1 Password Masking Versus Error Rate

Going into this study, we expected to see a lower error rate on unmasked passwords. This made sense logically—participants can visually inspect an unmasked password for errors, whereas masked passwords seem more prone to "fat finger" mistakes (touching an adjacent key without realizing it). Unmasked passwords did indeed have a lower error rate, but not nearly as dramatically as we would have expected–a 3% difference between masked and unmasked for a sample size of $n = 10$.

We plotted the data, looking for a difference between masked and unmasked passwords. There was definitely a pretty broad range in the time it took to enter a password, but masked and unmasked data points are nearly perfectly interleaved. We also compared the overall entry time for masked and unmasked passwords (Fig. 3, below), and while there was a difference, it was very small.

4.2 Device Type Versus Error Rate

It seemed clear that the difference between the error rates for masked and unmasked passwords was very small, so we started looking for other interesting features of the data. We had participants using both iPhone and Android devices—was one platform less prone to mistakes? (See Fig. 4).

Visually inspecting the plots, it was hard to see a difference between the Android and iOS columns.

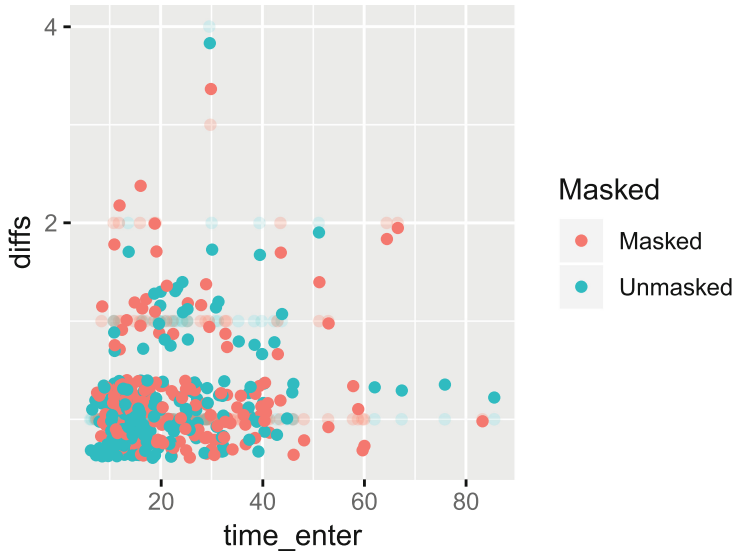

Fig. 2. Time to enter a password versus the number of differences between the expected and actual password. Jitter is added to the points in order to show the clustering of the data, but all differences are integer-valued.

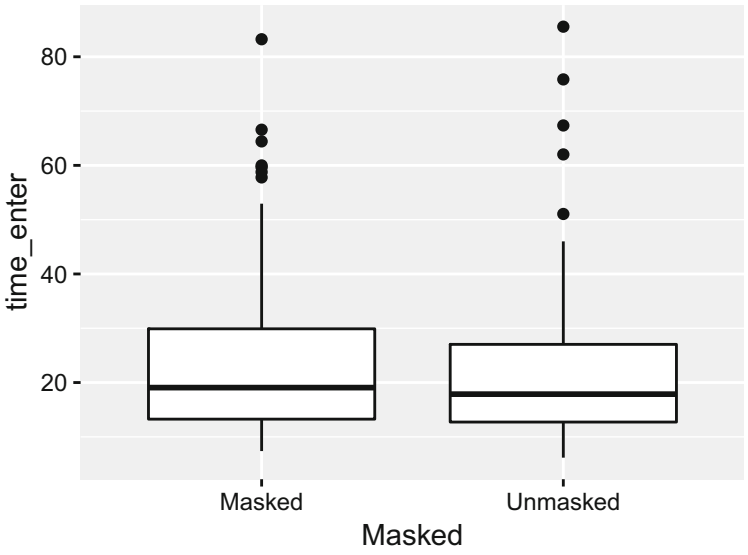

Fig. 3. Boxplot of the time it took to enter passwords with each password masking.

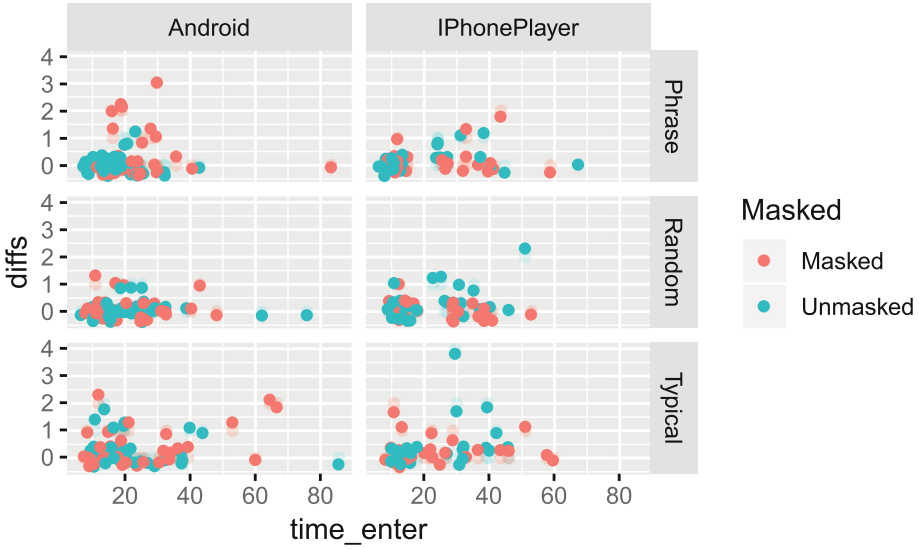

Fig. 4. Number of mistakes versus entry time, split by OS and password type. Points are jittered, y values are integers.

4.3 Password Type Versus Error Rate

Neither password masking nor device type affected the error rate more than is expected by chance. We were also curious about each of the password types—since a random password is much shorter than a multi-word password, we decided to compare the masked and unmasked entry times per password type. Given that all the passwords chosen for the study were equivalently secure, was a certain type of password more error prone? We plotted the same graph as shown in Fig. 2, this time coloring the points by the password type instead of whether it was masked or unmasked (See Fig. 5).

Visual inspection showed, once again, that there was no real difference, so we turned ourselves towards one last metric: was a certain password type faster to type than the others? (See Fig. 6).

Answer: not really. Even comparing each participant's typing speeds, the graphs for each password were more or less the same. This is actually interesting, given the difference in length between an 11-character random password and an upwards-of- 20-character multi-word password. This also aligns with similar research, which shows random passwords with special characters took considerably longer to type than standard text of the same length [5]. As for why this occurs, our qualitative results were in agreement with the previous work: users have to spend considerable time hunting for special characters.

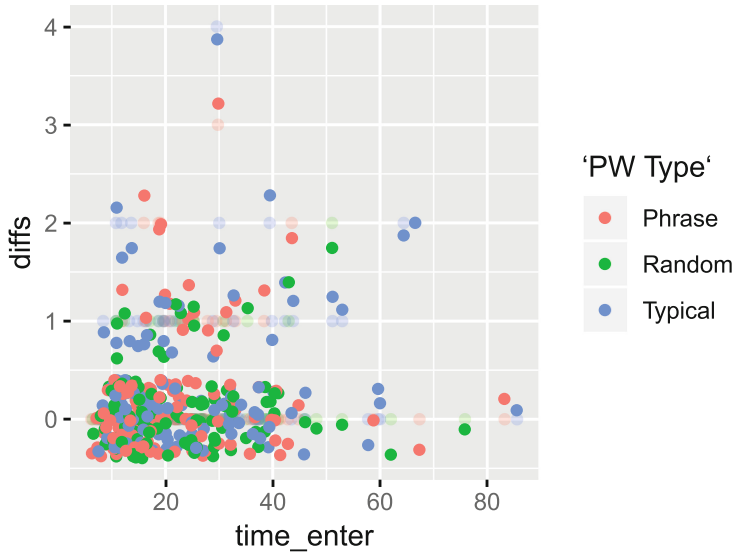

Fig. 5. Number of mistakes versus entry time, split by password type. Points are jittered, y values are integers.

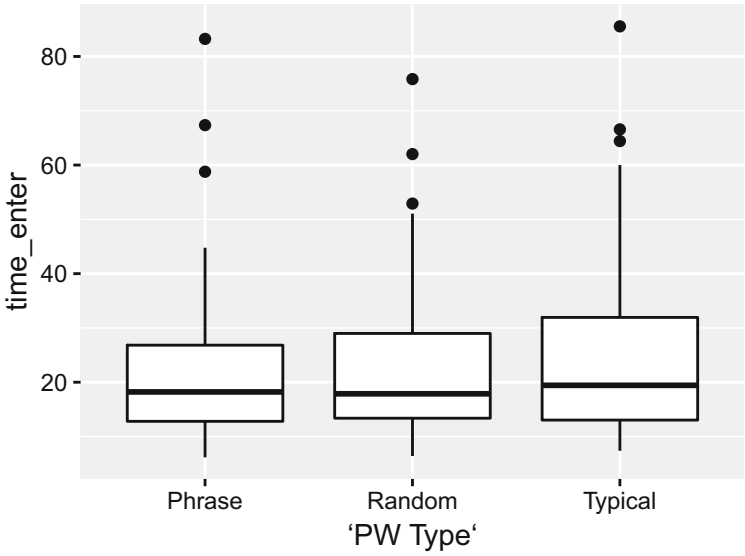

Fig. 6. Entry time versus password type.

4.4 Qualitative Analysis

Following each user session, we asked the same three exit questions.

Do you remember any of the passwords you typed? One participant could name nearly half the passwords, even some random ones, letter for letter. Another could not name a single thing they typed. The random passwords were definitely the least remembered, and while people could usually remember parts of the other two types, there were frequent mistakes. For example, they would confuse the order of words in a multiword password, or mix the words up between multiword passwords. For more typical passwords, they often remembered words, but would make mistakes on the exact placement of special characters, such as confusing "`Cr@ck3rJacked`" as "`Cr@ckerJack3d`".

Which of the password types, if any, did you feel were quicker to type? Nearly everyone said that the multiword passwords were easiest to type. We were surprised, however, that several participants also added, unprompted, that they would not choose such passwords since the lack of special characters and numbers made them less secure.

Which of these password types, if any, did you feel were easier or more difficult to type? For this question, we showed them a list of the passwords they had typed for the study, and asked them to point out any passwords that were particularly easy or difficult, and why. There was one password that was significantly longer than the other passwords, "`jubilant wineshop sceptic cadenza`", which was nearly always pointed out as difficult when participants encountered it. They would also point out characters they found ambiguous (such as a lower case 'l' and the upper case vowel, 'I').

5 Discussion

We observed a 3% higher error rate when passwords are masked versus unmasked. Based on the study size, we do not claim that this difference is real. We also could not find any groups or subgroups that had different error rates. If our data is at all typical, it seems that users are simply too good at entering passwords for it to matter whether or not they can see what they are typing, and thus we cannot reject the null hypothesis that masking or unmasking passwords makes no difference to the entry error rate.

6 Threats to Validity

This study was carried out as a small academic project, and as such, does not have a large time or financial budget. So while this project is intended to serve as a proof-of-concept for later work, it is not an exhaustive answer to the questions at hand.

6.1 Participants

The first threat is our sampling and number of participants. Our user study used a recruitment agency rather than surveying random students, which was done expressly to limit selection bias. While we had a good variety of ages, genders, and smartphone experience, candidates from this agency are often unemployed (which makes sense, given that the study took place on a weekday morning and afternoon). We also excluded UI/UX experts from our selection. Neither of these factors were a problem for the purposes of our work, but it does mean that the selection was not truly random. And even if selection were indeed random, the small number of participants ($n = 10$) limits the extensibility of our results.

It is also important to note only one participant was an English native speaker. While everyone was fluent enough to converse and they could use their keyboard of choice, both the "typical" and "multiword" passwords were based off of english words. This easily could have affected the entry time and error rate—unknown words take longer to visually parse, and are much easier to make mistakes when re-typing due to this lack of familiarity. This also could have made them more difficult to remember, thus hindering the learning response.

6.2 App

The app used for the study had some issues. One participant found a bug where pressing the back button would cause the keyboard to flicker on and off for several seconds, preventing the user from entering text. This did not happen a lot, but it still affected the password entry times. The font choice also proved to be problematic. We chose a monospace font with the goal that it would be easily readable, but one participant had considerable trouble with the glyph for the zero digit, and ended up looking for a special symbol that was an exact match to the font choice. Others asked about ambiguous characters, and still others just made their best guess and moved on.

The app was built specifically for this study. Therefore, it was not feasible to place the app directly onto participants' phones. Participants therefore had to use phones that were potentially unfamiliar to them. Typing on an unfamiliar device may lead to both more typing time and errors, and while we hoped to prevent this as much as possible, we could not completely remove the problem.

There is also the problem of the password selection used in the app. The random and multiword passwords used for the study were all generated by 1Password, and the typical passwords were generated by hand. The 1Password password strength metric was used to measure all three password types (to make sure they were roughly equivalent security-wise). However, 1Password is very clear that a strength metric can only test the strength of passwords for which it understands the underlying system that created them [6], meaning it could not accurately assess the security of our typical passwords.

6.3 Analysis

With $n = 10$, this study is clearly underpowered. Additionally, the number of ways we have looked at the data makes it very difficult to draw interesting conclusions, especially given the small sample size. Even if we did find a few interesting trends, we did so many other comparisons that the necessary corrections to any p-values would, in our opinion, render any hypothesis test dubious.

7 Conclusions

After all the debate over the tradeoff between the usability of unmasking passwords and the security of keeping them hidden, it seems that password entry is mostly working as-intended. There were slight differences in the error rates of masked passwords and their unmasked counterparts, but this difference was not statistically significant, nor were any of the other ways we sliced and diced the data. People are simply too good at typing passwords for it to matter whether they can see all the characters.

To replicate the research and the figures, download the raw data file (with participant names anonymised) and the R [16] scripts used to produce the figures from https://github.com/sten13/password-masking-data. The PasswordResearcher app is also available from https://github.com/catiejo/PasswordResearcher.

Acknowledgments. The authors wish to thank Bernhard Tellenbach of the Information Security group, Jürgen Spielberger of the Distributed Systems Group and Marc Rennhard of the Institute of Applied Information Systems at the Zurich University of Applied Sciences for funding this research.

References

1. Bonneau, J., Preibusch, S.: The password thicket: technical and market failures in human authentication on the web. In: Proceedings of the Workshop of Economics in Information Security (2010)
2. Chiasson, S., Forget, A., Biddle, R., van Oorschot, P.C.: User interface design affects security: patterns in click-based graphical passwords. Int. J. Inf. Secur. **8**(6), 387–398 (2009). https://doi.org/10.1007/s10207-009-0080-7
3. Chiasson, S., Oorschot, P.C.V., Biddle, R.: A usability study and critique of two password managers. In: Proceedings of the 15th USENIX Security Symposium, pp. 1–16 (2006)
4. Enge, E.: Mobile vs desktop usage in 2018: mobile widens the gap, April 2018. https://www.stonetemple.com/mobile-vs-desktop-usage-study/
5. Gallagher, M.A.: Modeling password entry on mobile devices: please check your password and try again. Ph.D. thesis, Rice University, May 2015
6. Goldberg, J.: Toward better master passwords, August 2018. https://blog.1password.com/toward-better-master-passwords/

7. Greene, K.K., Kelsey, J., Frankli, J.M.: Measuring the usability and security of permuted passwords on mobile platforms. Technical report NISTIR 8040, National Institute of Standards and Technology, Information Access Division, Information Technology Laboratory, 100 Bureau Drive (Mail Stop 8940) Gaithersburg, MD 20899-8940, April 2016
8. Gruschka, N., Iacono, L.L.: Password visualization beyond password masking. In: Bleimann, U., Dowland, P., Furnell, S., Schneider, O. (eds.) Proceedings of the Eighth International Network Conference (INC 2010), Heidelberg, Germany, 6–8 July 2010, pp. 179–188. University of Plymouth (2010). http://www.cscan.org/?page=openaccess&id=111
9. Gutmann, P.: Engineering Security (February 2014, in publication)
10. Herley, C., Oorschot, P.C.V.: A research agenda acknowledging the persistence of passwords. IEEE Secur. Priv. **10**(2), 28–36 (2012)
11. Holmes, J.: Stop password masking, September 2014. http://passwordmasking.com/
12. Melicher, W., et al.: Usability and security of text passwords on mobile devices. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI 2016, pp. 527–539. ACM, New York (2016). http://dx.doi.org/10.1145/2858036.2858384, http://doi.acm.org/10.1145/2858036.2858384
13. Mitchell, T.: Americans, password management and mobile security, January 2017. http://www.pewinternet.org/2017/01/26/2-password-management-and-mobile-security
14. Nielsen, J.: The need for web design standards, September 2004. https://www.nngroup.com/articles/the-need-for-web-design-standards/
15. Nielsen, J.: Stop password masking, June 2009. https://www.nngroup.com/articles/stop-password-masking/
16. R Core Team: R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria (2018). https://www.R-project.org/
17. Robinson, J.A., Liang, V.W., Chambers, J.A., MacKenzie, C.L.: Computer user verification using login string keystroke dynamics. Trans. Sys. Man Cyber. Part A **28**(2), 236–241 (1998). https://doi.org/10.1109/3468.661150
18. Schaub, F., Deyhle, R., Weber, M.: Password entry usability and shoulder surfing susceptibility on different smartphone platforms. In: Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia, MUM 2012, pp. 13:1–13:10. ACM, New York (2012). http://dx.doi.org/10.1145/2406367.2406384, http://doi.acm.org/10.1145/2406367.2406384
19. Schneier, B.: The pros and cons of password masking, July 2009. https://www.schneier.com/blog/archives/2009/07/the_pros_and_co.html
20. Wroblewski, L.: Mobile design details: hide/show passwords, November 2012. https://www.lukew.com/ff/entry.asp?1653
21. Yang, Y., Lindqvist, J., Oulasvirta, A.: Text entry method affects password security. In: The LASER Workshop: Learning from Authoritative Security Experiment Results (LASER 2014). USENIX Association, Arlington (2014). https://www.usenix.org/conference/laser2014/program/agenda/presentation/yang