Breaking: Password Entry is Fine

Abstract

In our digital world, we have become well acquainted with the login form — username shown as plaintext, password shown as 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 has been 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 things like typing speed, error rate, and number of backspaces. We then did an exploratory data analysis fo the data, and our findings show that, perhaps unexpectedly, there is no significant difference between masked and unmasked passwords for any of these metrics.

Introduction

The average person logs into between 7 and 25 accounts every single day [4], and the vast majority of people memorize their passwords, meaning reuse of identical or similar passwords is common [14]. This stands in opposition to traditional security advice, which dictates that passwords should be long, random strings full of complex characters that are 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 [14]. 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 it also takes 20% longer on average for them to do so [13].

Password unmasking is the practice of displaying passwords in plaintext rather than the traditional censored format. Is this a viable alternative? Usability experts claim that masking passwords is an unnecessary complication that causes confusion and frustration. Their more security-focused counterparts warn against the dangers of

shoulder surfing (stealing passwords by looking over people's shoulders as they type). While there are a number of articles and blog posts espousing one view or the other, there has been little academic rigor applied to 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 as it pertains to smartphone users?

To Mask or Not To Mask

Password masking dates back to the early days of terminal computing, when computers were mostly used by scientists, and every command was printed out on paper [10]. It made sense to replace passwords with asterisks at that point; without it, anyone with access to the computer printouts could easily harvest someone else's login info. But this rationale no longer applies. The days of needing to be an expert to use a computer are long gone, and there is no printout to worry about. Still, users and security experts alike balk at the idea of showing passwords in plaintext. What if someone looks over your shoulder without you 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 look away [19]? From a usability perspective, password masking also helps communicate that a password is sensitive information [19]. It is also what the user expects — a recent study on password unmasking showed that 60% of users grew suspicious when their passwords were unmasked by default [12]. Even though masking is a purely cosmetic fix, removing it tends to make users think there is a bug or even a security threat on the site in question [12].

On the side of password unmasking, prominent usability experts point out that the lack of visual feedback flies in the face of usability guidelines [19][16]. An error-prone login experience frustrates users, and it can even cause a security threat. Rather than risking having to retype a complex password, users opt for simpler versions, or even copy-pasting from a text file [19][16][14]. Additionally, it is much harder to snoop a password typed onto a smartphone screen, and with mobile internet usage recently overtaking desktop traffic [5], the threat of shoulder surfing is becoming less common relative to the nuisance of repeatedly typing a password [20].

Passwords are not dead. Despite the rise of biometrics, OAuth, and other forms of authentication, passwords still remain the most used on the internet today [11].

Therefore, it is important to continue to study usernames and passwords, despite their flaws, because of the sheer volume of people they affect. Similarly, mobile phones are not disappearing in the immediate future, so gaining a deeper understanding of entering a password on a mobile device is still a relevant issue. For as long as passwords have been around, we are still in a "data poor" research state [11], so this study contributes to password research as well as the growing intersection between usability and security research.

Contributions

We specifically make the following contributions:

- We compare the error rates of different types of password fields on mobile phones.
- We tested 2 different types of login screens: the standard, masked password, and a completely unmasked password.
- 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.

Related Work

In this section, we cover related research areas which have been divided into three main categories.

Usability of Password Masking

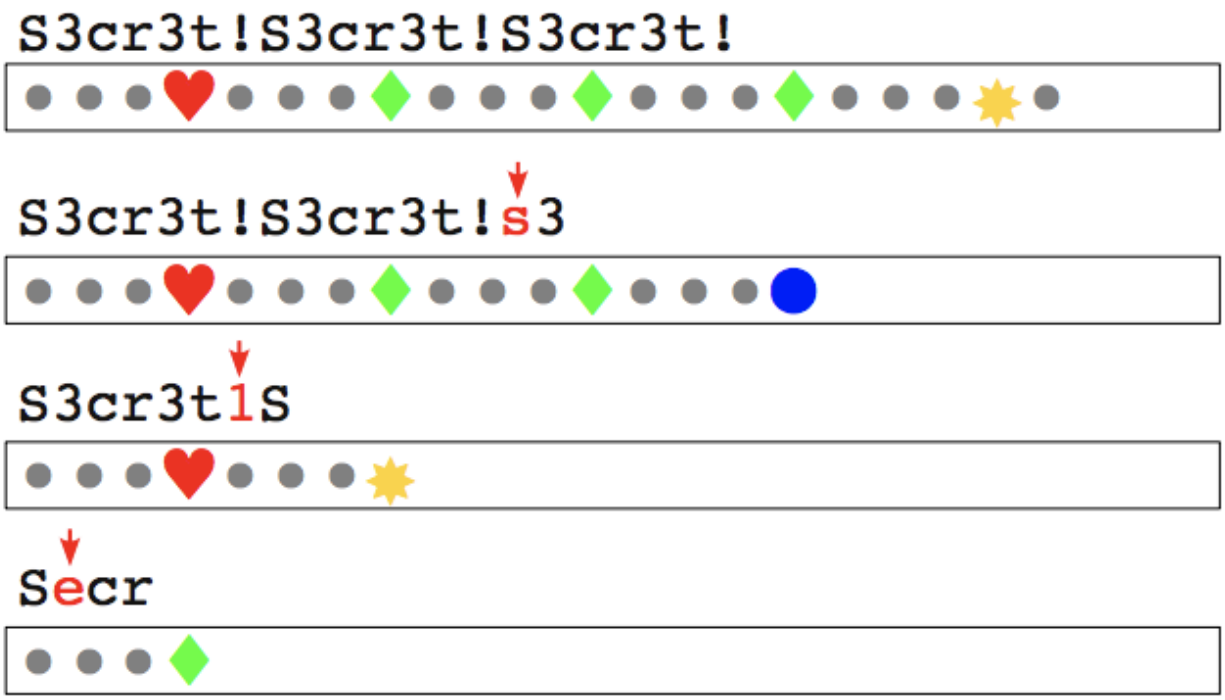
There has been considerable debate over the usability of masked passwords. Usability experts Jakob Nielsen [16] and Bruce Schneier [19] have argued that masking passwords causes users to choose easier (and therefore less secure) passwords, or equally problematic, copy and paste them from a file. They downplayed the threat shoulder surfing, a point which many of the commenters on these articles disagreed with. They argued that using a computer in a public setting like a coffee shop, as well as needing to enter a password during a presentation, are legitimate use cases for password masking. Additionally, a small-scale user study also showed that unmasking passwords broke users' expectations and even undermined their trust — 60% of participants became suspicious of the site when their passwords were displayed in plaintext [12].

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

there are good points on either side, how much of a difference does an unmasked password really make?

Usability & Security of Different Types of Logins

There is also 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 non-authorized users. However, typos and the subsequent use of backspace interfere with this metric, making it not feasible for large-scale adoption. Another study explored an alternative to password masking, the TransparentMask [9]. 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.



Transparent Mask example [9]

While these ideas are fascinating, their likelihood for widespread adoption is unlikely, and could also lead to user confusion. According to a survey of web tools done by Jakob Nielsen [15], 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][11].

Mobile Password Entry

As mobile phone usage continues to increase [5], passwords are increasingly being entered on mobile devices. And while mobile devices are often cited as a reason why password unmasking is important, we did not find studies measuring how much masked password fields and tiny keyboards affect usability. 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 types of mistakes commonly made when entering a password [6]. Prior research has also explored the error rate of typing passwords on mobile phones compared to desktops [13], 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 [8]. 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.

Methodology

This study aims to answer two questions. The first is how does password masking affect the error rate of password entry on mobile devices? Secondly, how does the makeup of of a password influence the aforementioned error rate?

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.

Effects of Password Makeup

While there is a lot of variation in passwords, 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 that used by a password manager ("nBqzEcP2A}Q8,jG")
- **Bad passwords**: passwords most commonly 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.

Study Structure

Our study involves the use of a custom-built app, PasswordResearcher, for user testing. 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 was 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 a given password three times, appearing in a random order. Entering the password multiple times is intended to elicit a learning response since participants learn to type their passwords more quickly and consistently over time. Additionally, the random shuffling is to allow 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 over time.

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 decided that while we want to avoid tricking our participants for ethical reasons, we also wanted to be careful not to mention that we are interested in the error rate of password entry since this could change their natural behavior and skew our results.

Data Collected

We used a recruiting agency, TestingTime, to recruit 10 participants. For each password attempt, we kept track of the following data:

- The expected password
- The actual password typed
- The time taken to enter the password
- The type of password (pass phrase, random, or typical password)
- Participant ID
- Overall attempt number (1 to 18 for each half)
- Password attempt number (1 to 3)
- Whether attempt is done on Android or iOS

It is worth noting that this study is only concerning itself with whether a login attempt would be successful or not, so we are only interested in 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 what we are testing.

We also collected qualitative exit questions from each participant:

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

Results & Analysis

We defined an incorrect password attempt as one where the actual password entered by the user does not match the expected password. Rather than simply using a boolean “correct/incorrect” however, we further quantified the correctness of a password by how many differences there were between the expected and actual password. We used the Levenshtein edit distance to compute the number of differences.

After collecting the data, we removed any outliers. There were some attempts that took an unrealistically long time (often because the participant set down the phone to come ask a question), so we threw out any attempts that took more than 100 seconds. Similarly, there were a few cases where participants accidentally pressed the enter key too soon, resulting in a high number of differences between the expected and actual passwords. To address this, we filtered out any attempts with more than 10 differences.

Because of the small sample size and study constraints, we opted to explore and visualize results without trying to prove statistical significance (see “threats to validity”). The conclusions below are our subjective interpretation of the data.

Password Masking versus Error Rate

Going into this study, we expected to see a lower error rate on unmasked passwords. This made logical sense — participants can visually inspect an unmasked password for errors, whereas masked passwords seemed more prone to “fat finger” mistakes (touching an adjacent key without realizing it). So we computed the raw 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. Unmasked passwords did indeed have a lower error rate, but not nearly as dramatically as we would have expected:

- Masked error rate: 0.1751412
- Unmasked error rate: 0.1542857
- Overall error rate: 0.1647727

We then started plotting the data, looking for visual indications of a difference between masked and unmasked passwords. Perhaps unmasked passwords were

quicker to type? After all, the participant could visually inspect the unmasked password for errors after typing it.

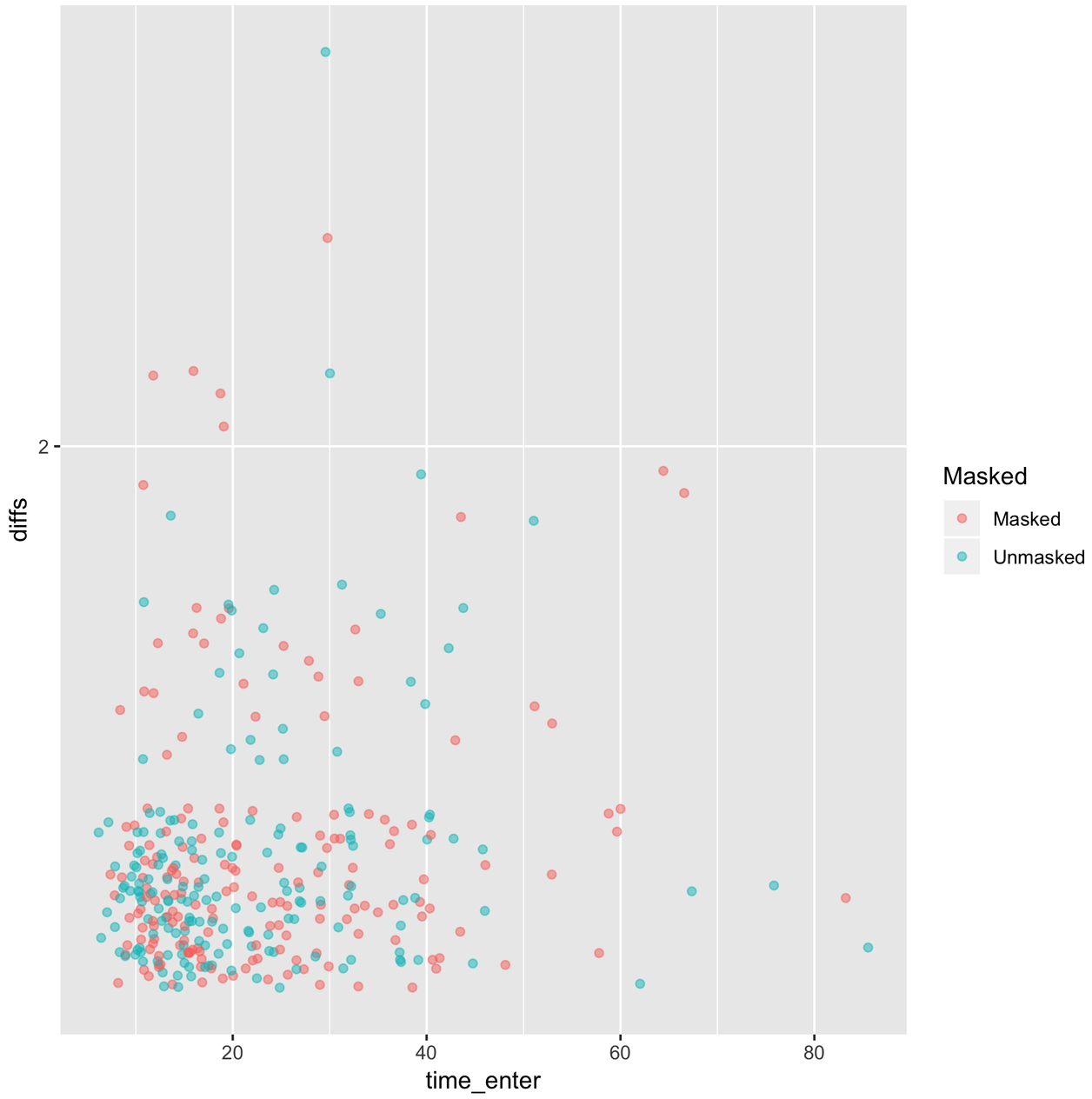


***Figure 1**: Time to enter a password versus the number of differences between the expected and actual password. Jitter is added to the points in order to show the clustering of the data, but all differences are integer-valued.*

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 next compared the overall entry time for masked and unmasked passwords (Figure 2, below), and while there was a difference, it was quite slight.

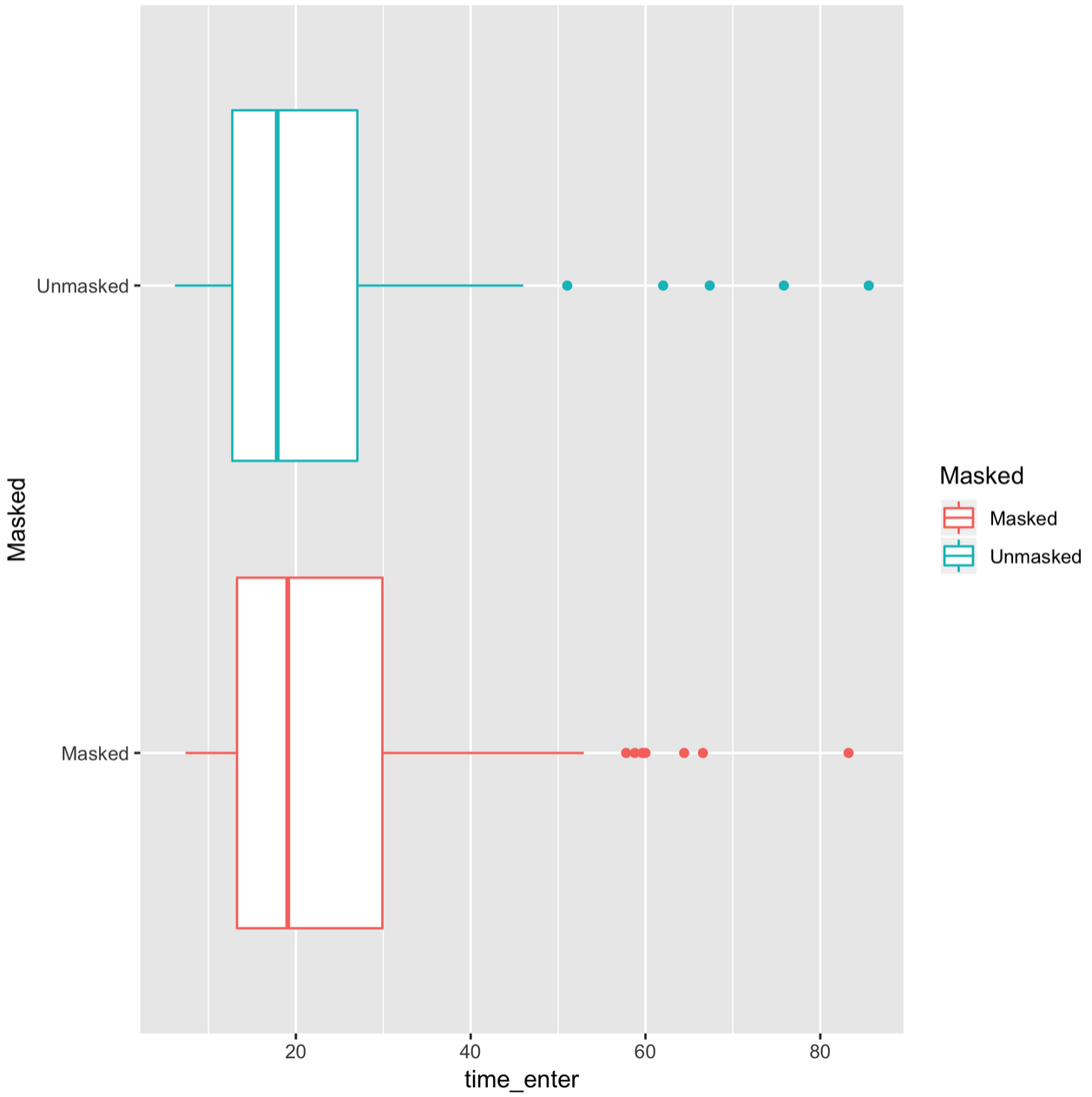


Figure 2: *boxplot of each the time it took to enter passwords with each password masking.*

Just to be sure, we also grouped the data based on the edit distance between the expected and actual password (so we could compare correct attempts to each other, and also incorrect attempts to each other). There are some noticeable differences for passwords with 2 differences (shown below, left), but the number of attempts with 2 or more errors are not big enough to be usable datasets, as shown by Figure 4 (below, right).

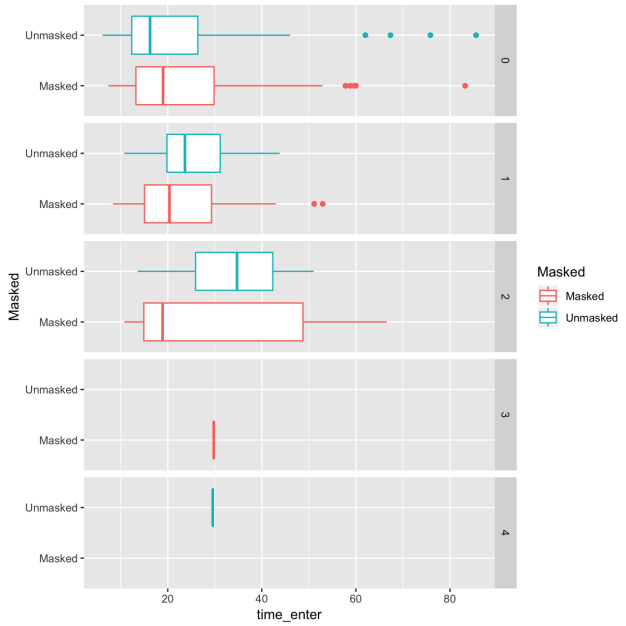


Figure 3: *same as Figure 2, but further grouping by the number of differences between the expected and actual passwords.*

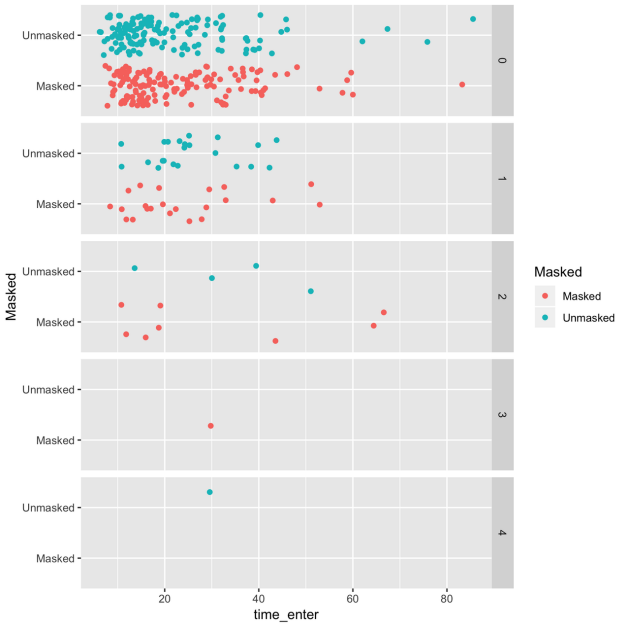


Figure 4: *Data points used to compute Figure 3, with jitter added to show data trends.*

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 easier to use and therefore less error prone?

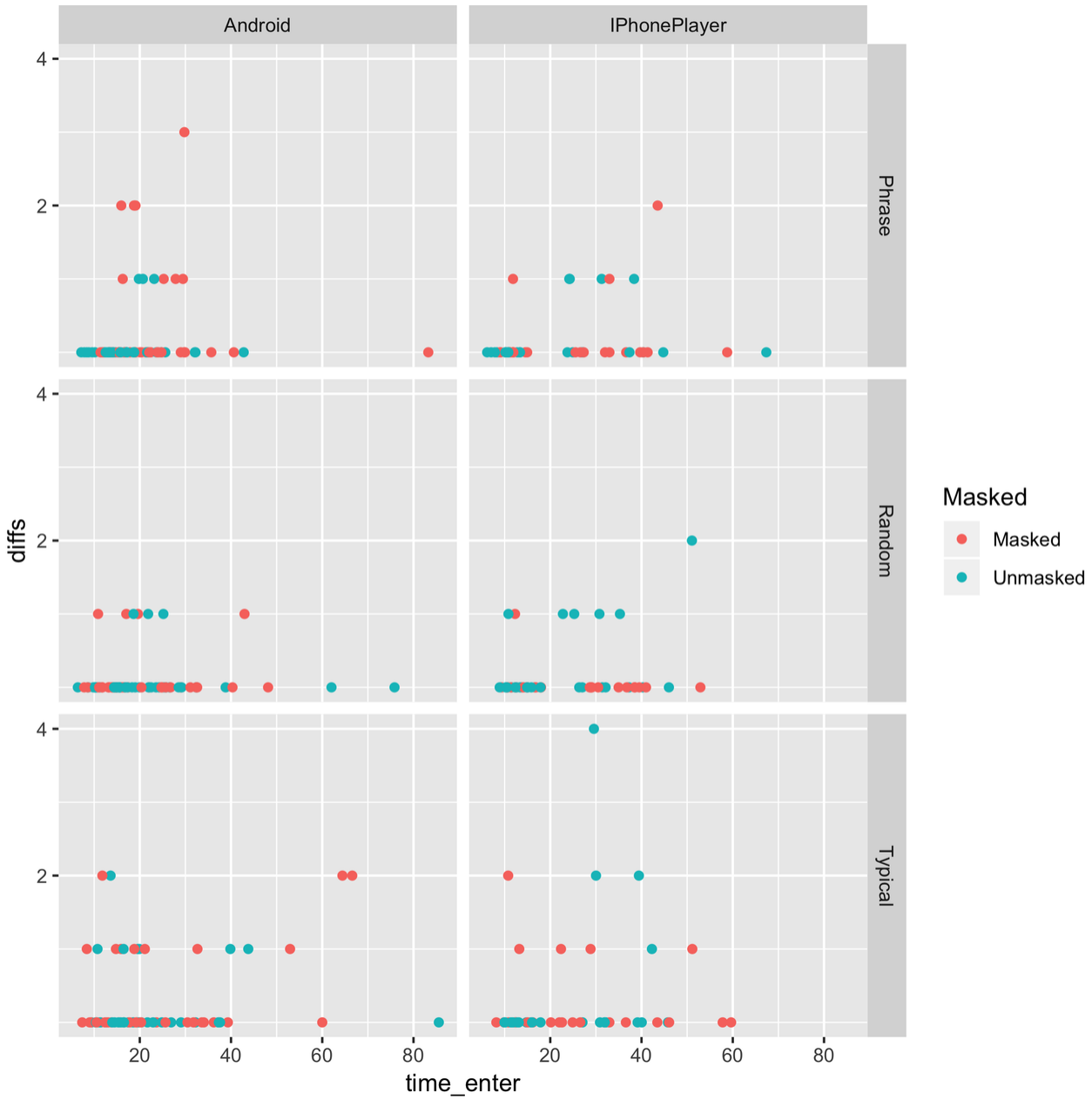


Figure 5: *the time to enter a password given the number of diffs, split by OS.*

Even just visually inspecting the plots, it was hard to see a difference between the Android and iOS columns. So we went back to the drawing board.

Password Type versus Error Rate

Neither password masking nor device type affected the error rate more than is expected by chance. Then we wondered 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 there a certain

type of password was more error prone? We plotted the same graph as shown in Figure 1, this time coloring the points by the password type instead of whether it was masked or unmasked.

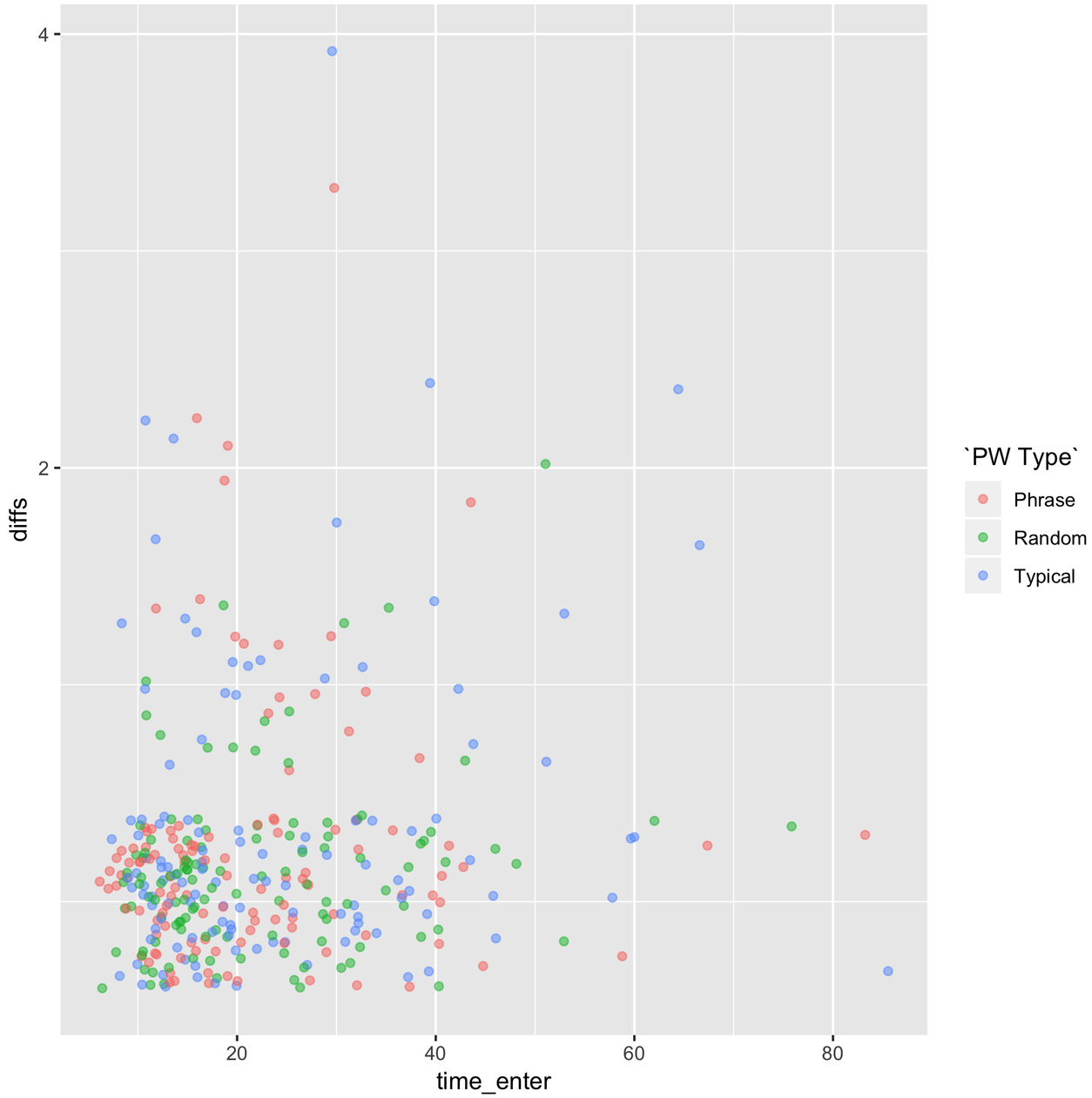


Figure 5: *differences between expected and actual password, colored by the password type.*

Just looking at the graph, it was clear that there was once again no real difference, so we turned ourselves towards one last metric: was a certain password type faster to type than the others?

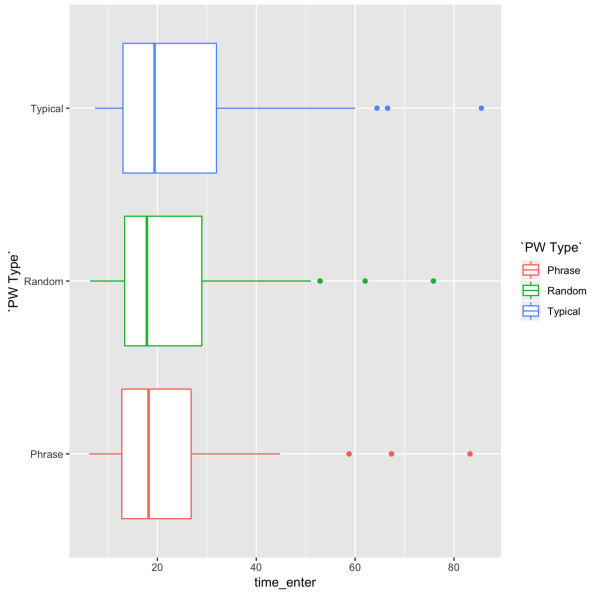


Figure 6: *entry time versus password type*

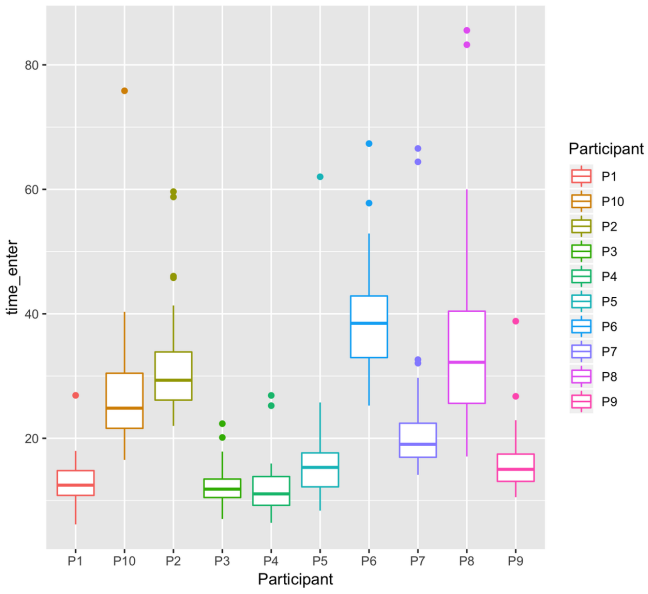


Figure 7: *entry time versus password type for each participant.*

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 [6]. 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.

Qualitative Analysis

Following each user session, we asked the same three exit questions. The findings of each of these questions are summarized in the sections below.

Do you remember any of the passwords you typed?

This was pretty varied. We had one participant who could name nearly half the passwords, even some random ones, letter for letter. Then there was another that 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 type of the password types, if any, did you feel were quicker to type?
Nearly everyone responded that the multiword passwords were easiest to type. We were surprised, however, that multiple participants also added, unprompted, that they would not choose passwords like this since the lack of special characters and numbers makes 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).

Final Conclusions

We observed a 2.1% higher error rate when passwords are masked. Based on the data size, we consider this to be insignificant. We also sliced and diced the data a number of other ways, and the answer was the clear: password entry is fine. And even if we had found interesting trends, the sheer number of ways we looked at the data would make it very hard to prove statistical significance. Despite all the debate that has gone on about password masking, it seems that users are simply too good at entering passwords for it to matter whether or not they can see what they're typing, and we must conclude that we cannot reject the null hypothesis.

Threats to Validity

As with all studies, there were factors in this experiment that could have affected to our results. It is important to note that this study is being 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.

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 the selection bias. However, while this is certainly better than asking friends and colleagues, it still cannot be completely free of bias. For example, we had a good variety of ages, genders, and smartphone experience, but 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 cannot be truly random. And even if selection was indeed random, we cannot ignore the fact that this study only looked at 10 participants. Neither our time nor financial budgets allowed for a larger group, but it is extremely difficult to get generalizable, statistically significant findings from such a small group.

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.

App

The app used for the study had a couple issues. There was a bug found by one of the participants 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 very many times, 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.

Since the app was built expressly for this study, it was not feasible to place the app directly onto participant’s phones. We had both Android and iPhone devices available, but this does not account for the variety of shapes and sizes of phones, nor

keyboard designs across different versions of Android. Typing on an unfamiliar device can 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 [7], meaning it could not accurately assess the security of our typical passwords.

Analysis

Finally, the number of ways we have looked at the data makes it very difficult to draw interesting conclusions, especially given the small sample size. We were unable to disprove the null hypothesis, so we also looked for other interesting correlations. Even if we did find a few interesting trends, we cannot do an honest analysis because we cannot just pretend we did not slice the data in all the ways that did not yield interesting results. If we look hard enough, we are bound to find something, so since we did not formulate specific hypotheses beforehand, we cannot in good faith relay these findings as scientifically valid.

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.

Works Cited

[1] Bonneau, J., & Preibusch, S. (2010, June). The Password Thicket: Technical and Market Failures

in Human Authentication on the Web. In *WEIS*.
[2] Chiasson, S., Forget, A., Biddle, R., & van Oorschot, P. C. (2009). User interface design affects
security: Patterns in click-based graphical passwords. *International Journal of Information Security*, *8*(6), 387.
[3] Chiasson, S., van Oorschot, P. C., & Biddle, R. (2006, August). A Usability Study and Critique of
Two Password Managers. In *USENIX Security Symposium* (pp. 1-16).
[4] Chisnell, D. (2011, September 22). *Random factoids I've encountered in authentication user
research so far*. Retrieved October 21, 2018, from http://usablyauthentical.blogspot.com/2011/09/random-factoids-ive-encountered-in.html
[5] Enge, E. (2018, April 27). Mobile vs Desktop Usage in 2018: Mobile widens the gap.
Retrieved from https://www.stonetemple.com/mobile-vs-desktop-usage-study/
[6] Gallagher, M. A. (2015). *Modeling password entry on mobile devices: please check your
password and try again* (Doctoral dissertation, Rice University).
[7] Goldberg, J. (2018, August 28). *Toward better Master Passwords*. Retrieved November 3, 2018
from https://blog.1password.com/toward-better-master-passwords/
[8] Greene, K. K., Franklin, J. M., Greene, K. K., & Kelsey, J. (2016). *Measuring the Usability and
Security of Permuted Passwords on Mobile Platforms*. US Department of Commerce, National Institute of Standards and Technology.
[9] Gruschka, N., & Iacono, L. L. (2010). Password Visualization beyond Password Masking. In *INC*
(pp. 179-188).
[10] Gutmann, P. (2014). *Engineering Security*. Auckland, New Zealand: University of Auckland.
[11] Herley, C., & Van Oorschot, P. (2012). A research agenda acknowledging the persistence of
passwords. IEEE Security & Privacy, 10(1), 28-36.
[12] Holmes, J. (2014, September 8). Stop Password Masking. Retrieved October 21, 2018, from
http://passwordmasking.com/

[13] Melicher, W., Kurilova, D., Segreti, S. M., Kalvani, P., Shay, R., Ur, B., ... & Mazurek, M. L. (2016,
May). Usability and security of text passwords on mobile devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 527-539). ACM.
[14] Mitchell, T. (2017, January 26). Americans, password management and mobile security.
Retrieved from http://www.pewinternet.org/2017/01/26/2-password-management-and-mobile-security/
[15] Nielsen, J. (2004). The need for Web design standards. *Recuperado el*, *14*.
[16] Nielsen, J. (2009). Stop password masking. *Retrieved October 17*, 2018.
[17] Robinson, J. A., Liang, V. W., Chambers, J. M., & MacKenzie, C. L. (1998). Computer user
verification using login string keystroke dynamics. IEEE transactions on systems, man, and cybernetics-part a: systems and humans, 28(2), 236-241.
[18] Schaub, F., Deyhle, R., & Weber, M. (2012, December). Password entry usability and shoulder
surfing susceptibility on different smartphone platforms. In *Proceedings of the 11th international conference on mobile and ubiquitous multimedia* (p. 13). ACM.
[19] Schneier, B. (2009, July 3). The Pros and Cons of Password Masking. Retrieved October 21,
2018, from
https://www.schneier.com/blog/archives/2009/07/the_pros_and_co.html
[20] Wroblewski, L. (2012, November 6). Mobile Design Details: Hide/Show Passwords. Retrieved
October 20, 2018, from https://www.lukew.com/ff/entry.asp?1653
[21] Yang, Y., Lindqvist, J., & Oulasvirta, A. (2014). Text entry method affects password security.
Proc. LASER, *154*.