Do Users Correctly Identify Password Strength?

Jason M. Pittman
High Point University
High Point, NC, USA
jpittman@highpoint.edu

Nikki Robinson
Capitol Technology University
Laurel, MD, USA
nrobinson@captechu.edu

***Abstract*—Much of the security for information systems rests upon passwords. Yet, the scale of password use is producing elevated levels of cognitive burden. Existing research has investigated the effects of this cognitive burden with a focus on weak versus strong passwords. However, the literature presupposes that users can meaningfully identify such. Further, there may be ethical implications of forcing users to identify password strength when they are unable to do so. Accordingly, the purpose of this study was to measure what socioeconomic characteristics, if any, led participants to identify weak and strong password strengths in a statistically significant manner. We gathered 436 participants using Amazon's Mechanical Turk platform and asked them to identify 50 passwords as either weak or strong. Then, we employed a Chi-square test of independence to measure the potential relationship between three socioeconomic characteristics (education, profession, technical skill) and the frequency of correct weak and strong password identification. The results show significant relationships across all variable combinations except for technical skill and strong passwords which revealed no relationship.**

Keywords—passwords, password strength, authentication, ethics, socioeconomic factors

I. Introduction

Password-based authentication is a prominent feature in modern life. Unfortunately, password authentication has grown to be an overwhelming burden to users [1]. In fact, Shay et al. [2] discovered the act of changing passwords on the premise of increasing password strength bothered users. Couple such results with the fact that users keep approximately 25 password- protected accounts [3][4], entering a password, on average, up to eight times each day, and one can imagine how sizable the growing cognitive epidemic may be.

The topic of conventional text-based passwords has been well studied [4][5][6]. In fact, there has been earnest effort to combat the inherent flaws in conventional password authentication through variations in form and recall modality [7][8][9][10]. However, the existing literature presupposes that users identify password strength accurately [11][12]. Indeed, Carnavalet [12] determined that the inconsistency in the password strength may be related to a misunderstanding of the characteristics required in a stronger password. This led us to wonder if an underlying motivation for such misunderstanding might be related to the users themselves. Accordingly, the purpose of this study was to measure what socioeconomic characteristics, if any, led participants to identify weak and strong password strengths in a statistically significant manner.

Furthermore, we considered not only password comprehension but also how ethics may be related to decisions to choose weaker passwords. That is, passwords strength is an ethical imperative from the perspective of an organization as many users work from home machines or workstations [13]. Later, research [14] found individuals are increasingly using more personal devices in the workplace because of the spreading trend of Bring Your Own Device (BYOD). Employees working on sensitive information have a duty to protect such information, and to have strong passwords on personal accounts. Thus, we arrive at the question of whether it is ethical for an individual to choose a weak password, even if the individual is under the impression the selected password is strong, when safeguarding sensitive information. To thoroughly investigate such an inquiry, we must first understand if users can reliably identify password strength.

II. Method

Broadly, we conjectured that subjects would be able to identify weak passwords consistently. Password characteristics such as length, capitalization, inclusion of alphanumerical and symbol characters serve as significant context clues. Further, we imagined subjects would not be able to consistently identify strong passwords, particularly when such were intermingled with weak passwords of similar length and combination. More technically, the goal of this correlational research was to determine if socioeconomic characteristics have measurable interactions with password identification and to what extent any such correlation is positive or negative. To that end, we operationalized subject education level, profession, and self-reported technical skill as socioeconomic variables on one hand and successful identification of weak and strong passwords as password identification variables on the other. Further, we imagined a single instrument as a means of collecting data to evaluate our hypotheses.

A. Instrumentation

We designed our data collection instrument in three sections. The first section held a standard informed consent, including opt-out procedures, and required affirmation of participation before continuing to the second section. We did not collect personally identifiable information. Instead, we coded and organized data with a simple integer index ranging between one and 436.

The instrument's second section held all the demographic questions. We asked subjects to self-report on age, gender, profession, technical skill level compared to others they knew, as well as how many passwords they used daily. The first three questions in this section served to collect categorical data for our socioeconomic variables. Further, we designed the last question as a screening mechanism insofar as we wanted to include only those individuals using at least one password daily.

Then, a third section held 50 passwords with a bounded response set of weak and strong which also served to collect categorical data. These passwords were randomly generated in two phases according to standardized definitions of weak and strong [15]. The first phase generated 100 weak passwords, parameterized as length of one to seven characters, the set of characters bounded to alphanumerics only, and any numerical characters positioned at the end of string. Concurrently, the first phase generated 100 strong passwords, parameterized as greater than 8 characters, the set of characters bounded to alphanumeric, punctuation, and special symbols, and the numeric or symbolic characters randomly placed within the string. The second phase consisted of removing any obvious weak (e.g., a single character or short, blatant sequence like 123) and then random selection of 25 passwords in each category into a randomly ordered list.

B. *Participants*

To achieve a suitable sample, we used Amazon's Mechanical Turk [16][17][18] to recruit individuals from a general population (scoped to Mechanical Turk users, biased towards the subset willing to take part in a questionnaire-based study, and being native or near-native English speakers) as opposed to a specific profession, age, or education category. According to Amazon, Mechanical Turk has a disparate and global user population of more than 500,000 people from over 190 countries. Thus, a more diverse and representative sample could be obtained by using Mechanical Turk instead of a traditional recruitment (e.g., email). Further, participants proactively responded to the work posting as opposed to us soliciting individuals.

The final sample size was 436 after we removed 11 participants data for being incomplete. We did not provide subjects with any instructional information about password strength. While not controlled for, we did ask participants to avoid (a) searching for password strength definitions; (b) use tools such as password strength checkers; (c) or any form of outside help. That said, we recognize a limitation in our protocol exists insofar as we did not control for any of these behaviors.

The sample was demographically distributed with respect to age, gender, profession and self-reported technical skill [19]. Only two participants were under the age of 18, making them members of a vulnerable population (more on this below). However, our IRB review, which included the potential for protected category participants, categorized the risk for harm as minimal given the anonymity of both our instrumentation and Mechanical Turk.

C. *Hypotheses*

We broadly conjectured participant education, profession, and technical skill would show a relationship with successful identification of weak and strong passwords. Disambiguated however, this general hypothesis turns into six discretely testable statements. That is, each variable-education, profession, and technical skill- manifested one hypothesis for successful identification of *weak* passwords and one hypothesis for successful identification of *strong* passwords.

III. Results

In total, our sample participants generated 21,800 discrete password identification trials. We ran a Chi-square test of independence against these data using pairs of variables in sequence with our stated hypotheses. Further, we analyzed the set of passwords in our instrumentation according to entropy measures and nominal strength compared to overall perception of each password by participants. While not associated with our primary focus, we felt at least a cursory description of these results may shed light on where participants correctly or incorrectly identified passwords as weak and strong.

A. *Education*

We examined education as a potentially related variable first. We compared the level of education (9 levels) and frequency of correctly and incorrectly identifying both weak and strong passwords (Table I). For weak password analysis, participants with some high school and individuals at the doctorate level identified correct passwords 56% and 57% of the time, respectively. The other education levels, High School, Some College, Trade School, Associate's Bachelor's, Master's, and a form of Professional degree, ranged from 68% to 79% able to correctly identify weak passwords. All education levels were able to correctly spot weak passwords 70% of the time. While it is interesting that the least amount of education (Some High School), and most (Doctorate), had the lowest average, the overall average displayed the ability for participants to find the weak password.

TABLE I. Frequency of identifying weak and strong passwords correctly and incorrectly

Education	**Weak Correct**	**Weak Incorrect**	**Strong Correct**	**Strong Incorrect**
Some High School	14	11	13	12
High School	439	186	305	320
Some College	962	338	644	656

Education	**Weak Correct**	**Weak Incorrect**	**Strong Correct**	**Strong Incorrect**
Trade School	257	68	149	176
Associate’s	544	206	374	376
Bachelor’s	3804	1246	2282	2768
Master’s	2011	564	1268	1307
Professional	204	96	157	143
Doctorate	130	95	116	109

Note: For *Weak* - [8] = 64.89, p = 5.10E-11, α = 0.05, critical value of 15.5.

For *Strong* - [8] = 33.71, p = 4.58002E-05 0.05, critical value of 15.5. We reject the null hypothesis in both cases.

To dig into strong password analysis, all education levels were within an 8- point percentage, between 45 and 52%, able to spot a strong password. The groups with the highest scores of 52% were Some High School and Professional degrees, with a Doctorate level education missing by one point, at 51%. Average ability to spot a strong password was at 49%, showing a major decline in ability to spot a strong password. Each group of education level was able to spot a weak password than a strong password more often.

B. *Professions*

The second variable we evaluated for a relationship was participants’ self- reported profession. As a variable, Profession showed a significant relationship with identifying weak passwords ([32] = 153.19, p value of 0.00, α of 0.05 and a critical value of 46.19). Consequently, we rejected the null hypothesis. The only professions to correctly identify a weak password less than an 80% of the time were Manufacturing - Electrical (53%) and Religious (71%). The professions which were able to identify weak passwords between an 80 and 89% were Agriculture, Education-K-12, Construction, Government, and Scientific. All other categories of professions were able to correctly find weak passwords over 90% of the time. The professions which were able to correctly identify weak passwords 100% of the time were Broadcasting, Legal, Mining, Publishing, and Retail. Of all professions, the average score for finding weak passwords was 90%.

We also evaluated participants’ profession against identification of strong passwords. We rejected the null hypothesis for this variable as well based on the Chi-square test of independence results ([32] = 66.11, p value of 0.0003, α of 0.05 and a critical value of 46.19).

Strong passwords saw a much lower result from weak passwords, with an average of all professions only able to identify them 40% of the time. The professions with scores lower than 40% were Broadcasting, Education-College, Finance, Legal, Manufacturing - Other, Religious, and Transportation. This was interesting because both Broadcasting and Legal professions found the weak passwords 100% of the time.

The only group which was not able to identify any strong passwords was Manufacturing - Electrical, which showed this group had the lowest identification for both weak and strong passwords combined. The group with the highest overall identification rate of 60% was Mining.

C. *Technical Skill*

Overall it seemed that the participants were able to accurately depict technical skill level when it came to perceiving weak passwords (Table II). The group which self-reported as much less skilled than their counterparts averaged 72% when identifying weak passwords. The less skilled group up to the much more skilled group averaged 91% ability to identify weak passwords. Of all the groups, participants averaged an 87% ability to find weak passwords correctly. Finally, we compared technical skill to successful identification of strong passwords. Unlike with previous tests, we found no relationship between these variables (Table II).

TABLE II. OBSERVED FREQUENCIES OF PARTICIPANT TECHNICAL SKILL IDENTIFYING PASSWORDS

Technical Skill	**Weak Correct**	**Weak Incorrect**	**Strong Correct**	**Strong Incorrect**
Much less skilled	16	6	8	20
Less skilled	355	23	271	401
Same skill	2119	201	1367	1913
More skill	4097	411	2585	3607
Much more skilled	1778	155	1077	1390

Note: For *Weak* - [4] = 14.95, p value of 0.005, α = 0.05, critical value of 9.5.

For *Strong* - [4] = 5.93, p value of 0.205, α = 0.05, critical value of 9.5. We reject the null hypothesis in both cases.

D. *Password Analysis*

After completing the Chi-square tests for independence, we wanted to more closely examine details associated with passwords used in the instrument. The goal was to develop a richer picture of what participants perceived based on their weak or strong selections and the entropy of associated passwords. We offer only descriptive analysis here of associations between data points; there was no attempt to infer causation.

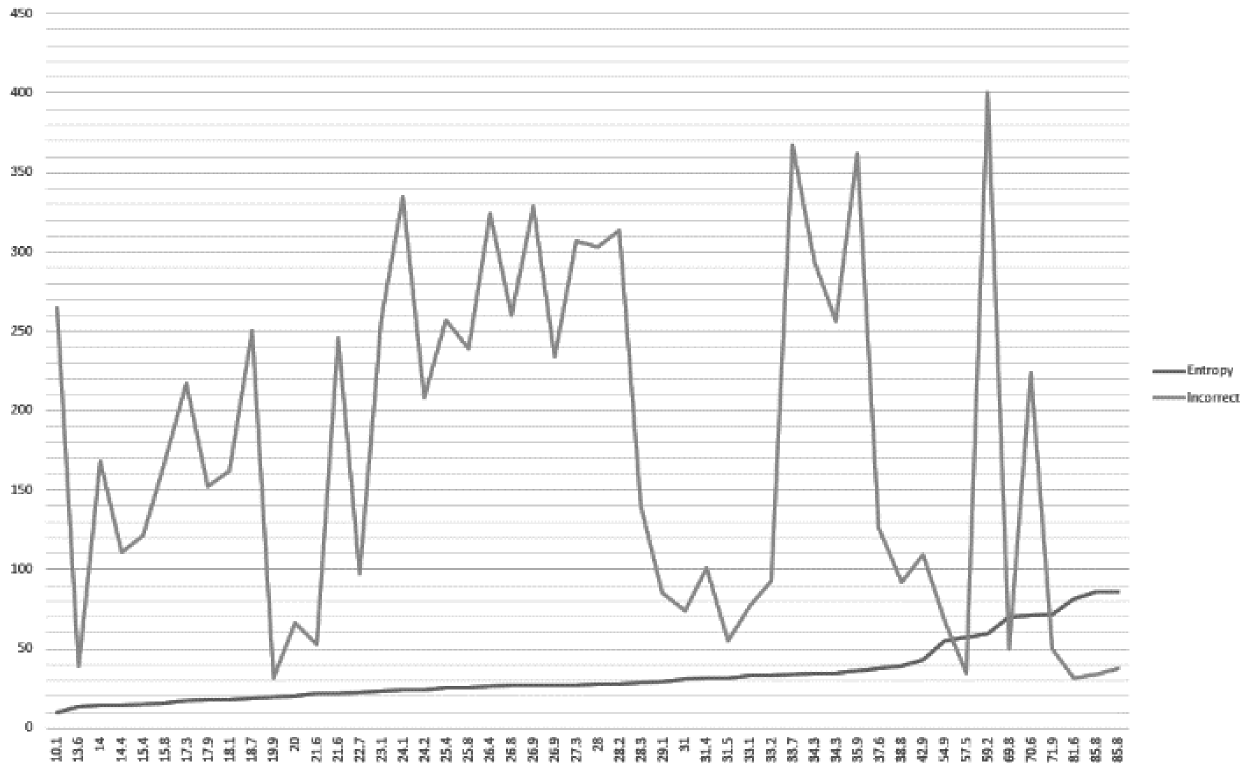

Fig. 1. Trend of incorrect perception of password strength as entropy in string rises

To begin, the most common incorrect choice was for the string G@m30f7hr0n3$. Four hundred participants incorrectly perceived this to be a weak password compared to 36 that correctly identified this as a strong password. The entropy is 59.2. Conversely, the most common correct choice was a tie between the string 1FcgiEF46Xy06jVS1 and qwerty. The former was a strong password while the latter was a weak password, and both were identified by 415 participants as such. The entropy of the two strings was 81.6 for the strong password and 19.9 for the weak. The biggest misconception for users was about strong passwords; overall, participants had a harder time identifying strong passwords. It would be interesting to find out why one password G@m30f7hr0n3$ was misidentified so often as a weak password. And if users felt this was such a weak password, why did they correctly identify 1FcgiEF46Xy06jVS1 as a strong password?

IV. Conclusions

The purpose of this work was to measure to what extent participant education, profession, or technical skill level are related to successful identification of weak and strong passwords. Towards this goal, we asked 436 human participants to judge whether 50 passwords were weak or strong. After data collection, we ran a Chi-square test for independence to measure relationships between variables and evaluate our hypotheses.

Education was significantly related to successful identification of weak and strong passwords alike. Further, each individual educational stratum showed higher frequencies of correct identification than incorrect. Based on these results, we can infer perception of what constitute a weak or strong password are not confined to any one educational stratum.

Profession was significantly related to successful identification of weak and strong passwords too. However, here we saw individual profession strata incorrectly identify password strength more often than correctly identifying password strength. While a stratum like Homemaker or Retired may not surprise anyone, the three Information Technology strata all more frequently misidentified strong passwords which is surprising.

There are a variety of follow up questions to be explored within the coupling of profession and perception of password strength. Experimental follow up may be of future interest to uncover what precisely causes specific professions to correctly identify weak passwords but incorrectly identify strong passwords. For a future study, we could focus specifically on individuals in IT fields, but target systems administrators, database administrators, and the like.

Interestingly, self-reported technical skill was significantly related to identifying weak passwords but not related to strong passwords. We wonder about the potential underlying factors contributing to this situation and emphatically suggest follow up research in this area. Because we saw some trending towards incorrect identification of strong passwords in various professions (e.g., Information Technology), we must wonder if such professions inherently harbor mentalities associated with incorrectly identifying strong passwords. Thus, we suggest any future work with technical skill not rely on self- reporting. Such study could robustly establish technical skill through empirical measurement.

Based on the inability to consistently identify weak and strong passwords, we wonder why individuals could not see clear patterns in password combinations. This could be due to the wide variety of password strength meters on websites, ranges in requirements for password strength, or even the sheer amount of different accounts users must create. Individuals will have a multitude of online accounts for business and personal reasons and would surely see conflicting information depending on the type of account. Without a dependable password format or template for all user accounts, individuals must use best guesses for creating secure passwords.

Participants without as much self-identified technical skill certainly had a more difficult time identifying weak passwords, which indicated that they are unable to create strong passwords in both business and personal accounts. Accordingly, a recommendation for these individuals is to evaluate their own passwords used on different accounts and create an exercise to change all passwords. This would allow the individual to explore current passwords versus a new set of stronger or more secure passwords. It would also provide an opportunity to use different passwords on all accounts, advancing their security awareness and education on password strength.

Based on these outcomes, we postulate that individuals with less self- identified technical skill will not have strong passwords in either business or personal user accounts. This leads to the question that if the participants do work in a professional setting, is there an absence of security training on strong passwords? Do the businesses these participants work in not have strong password requirements, or potentially unclear requirements? Superior security training and education should be made available to individuals in all industries. Tailored security guidance, specifically on password security and management, may enhance the ability to identify password strength. Along such lines, examples from this study could be used to support security training and explicitly demonstrate the differences between strong and weak passwords.

We would also strongly encourage that if users are still unable to identify weak passwords, to use multi-factor authentication (MFA) such as tokens or authentication applications to address this. MFA is a well-known tactic to protect against password attacks from malicious actors. Password attacks such as dictionary attacks, rainbow tables, and brute force attacks are common, and simple for a hacker to perform against any account. If security awareness training has not addressed password weakness issues, using MFA techniques would address security concerns and can be used in most business settings, and in personal accounts.

A few limitations were identified related to the conclusions. One limitation was the strength of the passwords; characteristics which define password strength are constantly evolving [20][21][22]. It is important to note that at the time of this study, the parameters for strong and weak passwords were based on current guidelines. And while the pool of passwords was quite large, the participants were limited to 100 passwords for the participants to evaluate. A final limitation was the self-reported information which was used to determine education, profession, and technical skill. The researchers chose these categories and options based on most relevant information in each field.

Lastly, we feel the ethical considerations of password selection ought to be investigated. While our findings demonstrate users' ability to correctly identify passwords, we are left wondering if users experience any ethical dilemma when selecting a new password. The dissonance between our results and the propensity for users to gravitate towards creating weak passwords hints at underlying factors worthy of inquiry. Similarly, we speculate there may be a discoverable balance related to how organizations force users to interact with passwords.

References

[1] De Joode, D. (2012). Does password fatigue increase the risk on a phishing attack? (Master's Thesis). Tilburg University, Tilburg, The Netherlands.

[2] Shay, R., Komanduri, S., Kelley, P. G., Leon, P. G., Mazurek, M. L., Bauer, L., Christin, N., Cranor, L. F. (2010). Proceedings of the Sixth Symposium on Usable Privacy and Security. doi: 10.1145/1837110.1837113

[3] Dhamija, R., & Dusseault, L. (2008). The seven flaws of identity management: Usability and security challenges. IEEE Security and Privacy, 6(2), 24-29. doi: 10.1109/MSP.2008.49

[4] Florencio, D. & Herley, C. (2007). A large-scale study of web password habits. Proceedings of the 16th international conference on World Wide Web, Banff, Canada, pp. 657-666. http://franklin.captechu.edu:2123/10.1145/1242572.1242661

[5] Notoatmodjo, G., & Thomborson, C. (2009). Passwords and perceptions. 7th Australasian Conference on Information Security.

[6] Komanduri, S., Shay, R., Cranor, L. F., Herley, C., & Schechter, S. (2014). Telepathwords: Preventing weak passwords by reading users' minds. 23rd USENIX Security Symposium.

[7] Brostoff, S. & Sasse, M. A. (2000). Are passfaces more usable than passwords? A field trial investigation. In Proceedings of HCI 2000.

[8] Jansen, W. A. (2003) Authenticating users on handheld devices. National Institute of Standards and Technology.

[9] Jermyn, I., Mayer, A., Monrose, F., Reiter, M., & Rubin, A. (1999). The Design and Analysis of Graphical Passwords. 8th USENIX Security Symposium.

[10] Wiedenbeck, S., Waters, J., Birget, J. C., Brodskiy, A., & Memon, N. (2005). Authentication using graphical passwords: Effects of tolerance and image choice. Proceedings of the 2005 Symposium on Usable Privacy and Security.

[11] Ur, B., Kelley, P. G., Komanduri, S., Lee, J., Maass, M., Mazurek, M. L., Passaro, T., Shay, R.,

[12] Carnavalet, X. C. & Mannan, M. From very weak to very strong: Analyzing password-strength meters. NDSS Conference. San Diego, CA.

[13] Dearman, D., Pierce, J. S. (2008). It's on my other computer!: Computing with multiple devices. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. doi: 1148/1357054.1357177

[14] Fleck, R., Cox, A. L., Robison, R. A.V. (2015). Balancing boundaries: Using multiple devices to manage work-life balance. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. doi: 10.1145/2702123.2702386

[15] Grassi et al. (2017). NIST Special Publication 800-63B, US Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD, https://doi.org/10.6028/NIST.SP.800-63b

[16] Turk, A. M. (2012). Amazon mechanical turk. Retrieved August, 17, 2012.

[17] Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. Judgment and Decision making, 5(5), 411-419.

[18] Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 453-456).

[19] Pittman, J. M., & Robinson, N. (2020). Shades of Perception-User Factors in Identifying Password Strength. arXiv preprint arXiv:2001.04930.

[20] Stavrou, E. (2017). A situation-aware user interface to assess users' ability to construct strong passwords: A conceptual architecture. International Conference on Cyber Situational Awareness, Data Analytics and Assessment. doi: 10.1109/CyberSA.2017.8073385

[21] Hart, D. (2015). Two studies on password memorability and perception, In 10th Annual Symposium on Information Assurance.

[22] Kankane, S., DiRusso, C., & Buckley, C. (2018). Can we nudge users toward better password management?: An initial study. In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (p. LBW593).