# Bars, Badges, and High Scores: On the Impact of Password Strength Visualizations

Maximilian Golla, Björn Hahn, Karsten Meyer zu Selhausen,
Henry Hosseini, and Markus Dürmuth
Horst Görtz Institute
Ruhr-University Bochum
{maximilian.golla, bjoern.hahn, karsten.meyerzuselhausen,
henry.hosseini, markus.duermuth}@rub.de

## ABSTRACT

Strength meters can help users to choose more secure passwords by representing strength via easy to understand textual and visual feedback. Bar-based meters representing strength as a progress supported by color and text are most frequently used. Non-bar meter visualizations are rarely studied and include radars, tachometers, and dancing bunnies. In this work, we consider alternative visualizations based on ideas that users often encounter in their daily lives. We explore gamification and peer-pressure as motivators, and test strength meters using badges and high scores based on a reward system similar to what typical video games offer. For a baseline, we consider a classical bar meter, as well as a control group without any strength meter. To evaluate the impact of these meters on the password strength, we performed a user study with 302 participants and a between-subjects design. Our findings support previous work, as no significant difference in password strength using various meter designs and motivators was found.

## 1. INTRODUCTION

User-chosen passwords are well-known for being comparatively easy to guess [20], thus password strength meters are well-studied in the academic literature [19] and widely used in practice [4]. A password strength meter (PSM) displays a representation of the estimation of the strength of a chosen password. It either tries to nudge or force a user to create a password that provides a reasonable level of security by means of guessing resistance. Widely adopted and intensively studied are meters that display the estimated strength as a *bar*, often horizontally oriented, typically with colors changing from red over yellow to green.

In this paper, we study meter variants based on game elements, in the form of high scores and badges. Gamification is well-known to be a powerful motivator for changing behavior [13]. We compare these to other meters based on peer-pressure [9], a traditional bar, and a condition without any strength meter. We conducted a study with 302 participants and a between-subjects design, collecting data for four different meter visualizations and one control group. In contrast to previous work [19, 9], we are the first to estimate password strength via the accurate *zxcvbn* [23] algorithm. Our findings support previous work, as no significant difference in password strength using various meter designs and motivators was found. In summary, our contributions are:

(i) We explore gamification as a motivator implemented as a high score and badge password strength meter.
(ii) We conducted a user study with 302 participants to evaluate the impact on password strength using an accurate strength estimation algorithm.
(iii) We compare our proposal with others, including a peer-pressure meter [9], a traditional bar meter, and a control group without any strength meter.

## 2. METHODOLOGY

Next, we motivate our meter choice, explain how we assess strength, and describe the tested conditions.

### 2.1 Motivators

Simply seeing the (estimated) strength in a visual representation results in a behavior change, i. e., users choosing stronger passwords [19]. Beyond simple *informative* meters, motivators such as *fear appeals* [21] and *peer-pressure* [9] have been studied. Egelman et al. explored the idea of using peer-pressure, which denotes the effect on an individual to change their behavior or attitudes, to conform with peers [9]. Their meter displayed a nudge that compares the users' password strength with those of peers: "Your new password is weaker than X % of users / stronger than Y % of users." However, they found no significant difference in the password strength.

In our work, we explore a motivator referred to as *gamification*, and describes "the use of game design elements in non-game contexts" [6, 5]. An introduction to the behavioral psychology of gamification is given by Walz and Deterding [22]. Gamification has been explored in the context of observing password and username creation [14], password strength perception [17], memorization [15] and implicit learning [1], security awareness and education [3], as well as for primary [8] and fallback authentication [16]. However, to the best of our knowledge, this motivator was never used in a


*Who Are You?! Adventures in Authentication (WAY)* 2018.
August 12, 2018, Baltimore, MD, USA.

password strength meter. We specifically use two elements from games. First, *high scores* are an important element in a large variety of games, where gamers can compare their performance with others. High scores can be based on items collected during the game, unlocked features, time to complete, or others. They encourage players to try harder and motivate top performance; they can motivate to continue playing a game or to replay games to improve one's skills. The high score is one game-based element that exhibits a strong connection with peer-pressure, as users are directly compared to peers. Second, many games offer a reward-based *badge system*, where different achievements during gameplay are honored with badges that are typically displayed in a "hall of fame." These badges can either unlock additional features during gameplay or offer no function at all and are entirely cosmetic and for prestige.

## 2.2 Strength Estimation

An essential aspect of a strength meter is the strength estimation. An accurate strength metric is important, as a faulty metric can influence users in the wrong way: If a weak password is rated strong a user might end up choosing this password, actually harming security, equally, if a strong password is rated weak the meter drives away the user from this strong password. We have chosen *zxcvbn* [23] due to its reasonably accurate strength estimates (cf. [11]). All strength estimates in our experiments are based on zxcvbn's guess number, where we took the logarithm to base 10 to obtain a more readable number from a smaller range. All meters use the same strength metric and the same minimum and maximum, in an attempt to make the results comparable.

## 2.3 Test Conditions

In our study, we tested five conditions. Two meters (*high score*, *badge*) that utilize game-based elements. The third, the *peer pressure meter*, is an adaption of the Egelman et al. [9] meter, with the primary intention to compare the *high score meter* with the *peer-pressure meter*, as both ultimately base on the idea of peer-pressure. We added two baselines: One straightforward *bar-based meter* and *no meter*, a simple password entry without any strength meter. Across all conditions, we use zxcvbn's guess number to estimate the strength and a sample of RockYou [2] passwords to calibrate the meter visualizations. Screenshots of all designs can be found in the Appendix A.

*High Score Meter.* The meter consists of two columns showing a rank and score value. A textual description "Y of X users have chosen stronger passwords than you!" is shown to the user. In the study, the high score does not display strength scores derived from passwords of other participants (as assumed by the user) but is filled in advance with values derived from 522 randomly sampled passwords from the RockYou leak. We have chosen 522 as we considered this number to be random-looking and reasonable high to convince the participants in our study to believe this represents the number of accounts that already registered with the (simulated) service. To display reasonable looking scores, we calculate $log_{10}(guess\ number) \cdot 100$.

*Badges Meter.* We designed nine different badges that a user can achieve by entering a stronger password. Badges are from three different categories: *length*, *strength*, and *blacklist*. For each category, there are three levels. The length category awards a badge if the length of the entered password is greater or equal than 6/8/10 characters. The strength category is based on the guess number of the password and changes its state by reaching thresholds of 33 %, 66 %, and 100 %. The blacklist category is based on three differently sized blacklists containing the 1000, 10 000, and 100 000 most common passwords in the RockYou leak. Active badges are colored and solid, and inactive badges are semi-transparent and displayed in grayscale. All nine badges include a label that describes the purpose of the badge.

*Peer-Pressure Meter.* We re-implemented the proposal from Egelman et al. [9]. It consists of a green and a red area. If a strong password is entered, the green area increases, while the red area decreases. Furthermore, the text "Your new password is - weaker than X % of users - stronger than Y % of users" is shown. The original and our reimplementation use the RockYou leak to calibrate the visualization. However, one cannot directly compare the results, as Egelman et al. filtered the leak using a different password composition policy, and afterward artificially inflated the derived thresholds. Furthermore, the implementations differ in the way they estimate strength. While Egelman et al. used a combination of the password length and keyspace, we use *zxcvbn* to estimate a password's guess number.

*Bar Meter.* The first baseline represents a generic bar-based strength meter. Our implementation follows the baseline from Egelman et al. [9]. The meter fills from left to right in 10 % steps and changes the color from red ($0 - 30$ %) over yellow ($40 - 60$ %) to green ($70 - 100$ %). At the same time, the visualization is supported by a label showing *Weak*, *Medium*, or *Strong*.

*No Meter.* The second baseline does not provide any meter and is indented to act as a control group. Participants in this group have not seen any strength meter during the account creation process. We included this condition to support the hypothesis that users who have access to any strength meter create stronger passwords than those who don't.

# 3. USER STUDY

Next, we describe our study protocol and sampling method.

## 3.1 Study Protocol

The design is adapted from previous work [9]. The participants were given a scenario where they had to create a new account for a university portal. (The real intention of the study was revealed in the debriefing.) The study consisted of six steps.

*In a first step*, we introduced the participants to *the portal* and described it to be indented for updating personal information, reading emails, viewing grades, and accessing the eLearning services.

*In a second step*, we asked to *create an account* with the portal to try out its functionalities. We randomly assigned each participant to one of the five test conditions. To increase validity, we asked participants for their email address, as we considered it to worth protecting. We displayed a blue colored label showing "How to make strong passwords." By hovering over the label, a message was revealed that contained three hints [12]. To prevent typos in the password entry form, we asked to confirm the password. We displayed

a composition policy "Use at least 6 characters" colored in red, which turned green at the moment the requirement was satisfied [18]. We enabled participants to display their password in plaintext by hovering over an icon.

*In the third step*, we presented a *questionnaire*, to distract the participant after creating their account. We asked participants how many hours per week they spend playing video games. The next questions differed by the randomly assigned condition. Participants that were assigned into one of the conditions that used a strength meter were shown a screenshot of their respective meter and the following questions (the *no meter* participants didn't see the questions): "You may have noticed a so-called *password strength meter* on the previous page. Such a meter estimates the resistance of your password against guessing attacks." In the following, we asked participants about the influence, the appropriateness, and their feeling on deploying the meter on their most visited website.

*In the fourth step*, we collected *demographic information* such as age, gender, and familiarity with computers. To prevent data collection errors, we also asked whether a participant already participated in the study sometimes before. The full questionnaire, including the demographics, can be found in the Appendix B.

*In the fifth step*, we *debriefed participants*, informing them about the real purpose of the study and asked them to recall their password. We used this as a sanity check to identify participants that entered a (random) password during account creation. After a successful recall, or 3 failed attempts participants were forwarded to the next step.

*In the sixth step*, participants were thanked for their help and asked to leave any form of *additional feedback*.

## 3.2 Participation and Sampling

We recruited 342 participants over a time span of 4 days in the central coffee lounge on campus. We had set up four computers in a relatively quiet corner of the lounge, and we had four persons supervising participants. This allowed us to collect a relatively large number of samples (needed to test five different conditions) with moderate effort. Also, this enabled us to recruit participants from all major faculties (engineering, humanities, medicine, and science), with the drawback of mostly recruiting students. We refrained from using an online study as we believe that the perceived value of an account can be higher in a local setting.

We didn't count the number of participants who exited early and abort the study, but we estimate the number to be in the range of 5 to 10 participants. Regardless of a successful participation or early abort, instructors compensated all participants for their time with some chocolate. On average completing the study took 4 minutes. Our sample represents a typical university population. The average age is 22 years but ranged from 18 to 56 years. 38 % of participants reported being female, 58 % identified as being male. Approximately, 22 % of our population reported to have a job, hold a degree, or major in computer science, computer engineering, information technology, or a related field. The full demographics can be found in the Appendix B.

Due to random sampling over four days and participants exiting the study early, the number of samples per condition is not the same: We collected 63 data points for *no meter*, 55 for *bar meter*, 65 for *peer pressure meter*, 62 for the *high score meter*, and 57 for the *badges meter*. 40 of the 342 participants reported having participated in the study before or knew the true goal of the study. Those participants were compensated as well but were excluded from the further analysis resulting in 302 participants.

## 3.3 Ethical Considerations

Users were informed that they were to take part in a scientific study and that no personally identifying information will be stored. We explicitly told all participants that we may ask for authentication information such as an email address and a password, but that we do not store this information. We informed participants that they can withdraw and stop participating at any time and that all partial data will not be analyzed or stored. Our institute does not have an ethics board or IRB, but we discussed the study design with peers to validate the ethical perspective of our research. We made sure to only store information which does not allow to link individuals to their collected data. Instead of storing a participant's password, we stored the respective strength value in an encrypted form. The email address of a participant was only displayed during the study but not stored on disk. While all questions were mandatory to answer, all demographic questions offered a "Prefer not to answer" option. We analyzed the (optional and anonymous) feedback of the study but did encounter any ethical issues or concerns.

# 4. RESULTS

All statistical tests use a significance level of $\alpha = 0.05$. We ran omnibus tests across all conditions, for each of the considered variables *password strength*, *time*, *user preference*, and *edits*. As none of the variables seems to follow a normal distribution, we used Kruskal-Wallis tests for the omnibus tests and Dunn's test for post-hoc analysis. The main results are shown in Table 1.

## 4.1 Password Strength

The collected strength scores, measured by the logarithm to base 10 of the *zxcvbn* guess numbers, are not normally distributed (Shapiro-Wilk normality test, $W = 0.96312$, p-value $= 6.05e-07$). We found no significant influence of the different meter visualizations on password strength (Kruskal-Wallis test, chi-squared $= 2.1308$, df $= 4$, p-value $= 0.7117$). We would have expected that at least the *no meter* condition leads to weaker passwords. While both median and mean of the strength are actually (slightly) lower for this condition, the differences are not significant. The results are compatible with previous findings: Ur et al. [19] found significant differences mainly for password length and very stringent meters, while the pure presence of meters did not change results. Previous work [19, 9] used different forms of strength estimation based on the theoretical keyspace, the length of a password, and a blacklist, which is not accurate and might explain the different results.

We analyzed the percentage of passwords that fall into the $10^6$ (online threshold) to about $10^{14}$ (offline threshold) classes, also know as *online-offline gap*, introduced by Florêncio et al. [10]. According to Florêncio et al. passwords that resist guessing attacks beyond a certain point can be considered a waste of cognitive effort since they deny an attacker nothing. Our numbers show that approximately 25 % of the users did

Table 1: Results Overview

| Test Conditions | Median Strength $log_{10}$(Guess No.) | Mean Strength $log_{10}$(Guess No.) | Median Time Seconds | Questionnaire Usage on Website | Median Edits No. |
|---|---|---|---|---|---|
| High Score Meter | 7.94 | 8.12 | 47.50 | Very satisfied (52 %) | 11 |
| Badges Meter | 8.00 | 8.35 | 42.00 | Slightly satisfied (26 %) | 11 |
| Peer-Pressure Meter | 8.00 | 7.83 | 38.00 | Moderately satisfied (28 %) | 11 |
| Bar Meter | 8.00 | 8.35 | 38.00 | Moderately satisfied (33 %) | 11 |
| No Meter | 7.81 | 7.39 | 38.00 | - | 10 |

create a password that could be at risk in an online guessing attack. We observed that only 5 % of the users created a secure enough password that could resist a severe offline guessing attack. Most importantly, approximately 70 % of the users fall into the online-offline gap. Thus, created a password that is good enough to withstand online guessing attacks, but too weak to resist an offline attack. These findings reject our hypothesis that game-based elements may provide effective strength meters, and seems to indicate that just any password meter will do.

## 4.2 Password Creation Time

We measured the time required to read the instruction ("*Before you can start, you have to create an account! Enter your email address and choose a strong password to continue.*"), enter the email address, type the password, and confirm it once. This variable is not normally distributed (Shapiro-Wilk test, W = 0.8356, p-value $< 2.2e-16$). A Kruskal-Wallis test found a significant influence of the conditions on the time for password creation (chi-squared = 15.734, df = 4, p-value = 0.003398). Post-hoc analysis (Dunn's-test for multiple comparisons, Bonferroni-corrected) showed significant differences between the high score meter and the peer-pressure meter (p = 0.0033), and between the high score meter and the no meter condition (p = 0.0281). The median time spent with the described process including the high score meter was 47.5 seconds, while for the peer-pressure meter and the no meter condition it was 38 seconds.

## 4.3 User Preference

We asked participants how they would feel if the respective meter would be used on their most visited website. For the high score meter, 54 % reported being very satisfied or better. For the bar and peer-pressure meter, the majority reported being moderately satisfied. For the badges meter, the feelings were mixed. While 26 % report that they would be slightly satisfied, 25 % report they would be very satisfied. A Kruskal-Wallis test (chi-squared = 1.4362, df = 3, p-value = 0.6971) found no significant difference in the answers.

## 4.4 Password Edits, Hints, and Recall Rate

We counted the number of key presses during password entry. We observed a median of 11 edits per password and meter. A Kruskal-Wallis test (chi-squared = 2.8335, df = 4, p-value = 0.5861) found no significant difference in the number of edits across the tested conditions. We measured whether the hint text "*How to make strong passwords*" has been read. We found that only 13 of the 302 participants read the text. The majority of 285 (94 %) were able to recall their password immediately. Only 7 participants (HS: 1, Bad.: 2, Peer: 1, Bar: 2, No: 1) were unable to authenticate within three attempts.

## 4.5 Questionnaire and User Feedback

Around 27 % reported not to play, while 20 % said to spent about $3-8$ hours on playing video games per week. Further, we asked participants how much influence the respective password strength meter had on their password choice. Participants reported a moderate influence by the high score and bar meter. No effect was reported by the participants for the badges and peer-pressure meter. The majority of participants agreed that their respective meter visualized password strength appropriately.

By analyzing the (optional) feedback, we got some more insights. For the high score meter, participants suggested changing the highlighting color with an increasing rank in the high score. One participant expressed concerns that the text ("Y of X users have chosen a stronger password than you!") might unintentionally leak security sensitive information. From this, we conclude that the design of the fictive high score entries, in fact, convinced some of the users. As mentioned before, the design of the badges meter did not match everybody's preference.

## 4.6 Limitations

Although the user study was planned and conducted carefully, it suffers from limitations related to the sampling and design. While the sampling approach allowed to reach for a large number of participants, it resulted in a distinct bias toward highly educated young participants. Albeit we used a quiet area, the environment in which the study was conducted was busier than a lab room. While we tried to come up with a setup similar to the one used by Egelman et al. [9], a limiting factor might have been that not all users were enough convinced from the setup to create a strong password. Finally, previous work successfully tested half score meters [19], for which users need to work harder to reach the same security level in the visualization. We would be interested to see whether a stricter meter would have led to different results.

# 5. CONCLUSION

In this work, we studied the impact of gamification and peer-pressure motivators in strength meters. We tested four different meters and a baseline using no meter, in a user study with 302 participants. We found very few differences between the meters, specifically no significant differences in the resulting strength of the passwords. We observed significantly longer password creation times using our high score meter design. Our results indicate that even drastically different visualizations have little to no effect on the strength of passwords and that non-standard meters may have an adverse effect on the creation times and thus the usability of password strength meters.

# 6. REFERENCES


[1] H. Bojinov, D. Sanchez, P. Reber, D. Boneh, and P. Lincoln. Neuroscience Meets Cryptography: Designing Crypto Primitives Secure Against Rubber Hose Attacks. In *USENIX Security Symposium*, SSYM '12, pages 129–141, Bellevue, Washington, USA, Aug. 2012. USENIX.

[2] N. Cubrilovic. RockYou Hack: From Bad To Worse, Dec. 2009. `https://techcrunch.com/2009/12/14/rockyou-hack-security-myspace-facebook-passwords/` , as of June 25, 2018.

[3] A. Dabrowski, M. Kammerstetter, E. Thamm, E. Weippl, and W. Kastner. Leveraging Competitive Gamification for Sustainable Fun and Profit in Security Education. In *Summit on Gaming, Games, and Gamification in Security Education*, 3GSE '15, Washington, D.C., USA, Aug. 2015. USENIX.

[4] X. de Carné de Carnavalet and M. Mannan. From Very Weak to Very Strong: Analyzing Password-Strength Meters. In *Symposium on Network and Distributed System Security*, NDSS '14, San Diego, California, USA, Feb. 2014. The Internet Society.

[5] S. Deterding, D. Dixon, R. Khaled, and L. Nacke. From Game Design Elements to Gamefulness: Defining "Gamification". In *International Academic MindTrek Conference: Envisioning Future Media Environments*, MindTrek '11, pages 9–15, Tampere, Finland, Sept. 2011. ACM.

[6] S. Deterding, R. Khaled, L. E. Nacke, and D. Dixon. Gamification: Toward a Definition. In *CHI Workshop on Gamification*, CHI '11, pages 6–9, Vancouver, British Columbia, Canada, May 2011. ACM.

[7] K. Dycha. Fantasy Icon Pack by Ravenmore, June 2012. `https://opengameart.org/content/fantasy-icon-pack-by-ravenmore-0`, as of June 25, 2018.

[8] F. Ebbers and P. Brune. The Authentication Game – Secure User Authentication by Gamification? In *Conference on Advanced Information Systems Engineering*, CAiSE '16, pages 101–115, Ljubljana, Slovenia, June 2016. Springer.

[9] S. Egelman, A. Sotirakopoulos, I. Muslukhov, K. Beznosov, and C. Herley. Does My Password Go Up to Eleven?: The Impact of Password Meters on Password Selection. In *ACM Conference on Human Factors in Computing Systems*, CHI '13, pages 2379–2388, Paris, France, Apr. 2013. ACM.

[10] D. Florêncio, C. Herley, and P. C. Van Oorschot. Pushing on String: The "Don't Care" Region of Password Strength. *Communications of the ACM*, 59(11):66–74, Oct. 2016.

[11] M. Golla and M. Dürmuth. On the Accuracy of Password Strength Meters. Manuscript, 2018.

[12] Google. Account Help – Follow Tips for a Good Password, Dec. 2017. `https://support.google.com/accounts/answer/32040?hl=en`, as of June 25, 2018.

[13] J. Hamari. Do Badges Increase User Activity? A Field Experiment on the Effects of Gamification. *Computers in Human Behavior*, 71(C):469–478, June 2017.

[14] D. R. Lamichhane and J. C. Read. Investigating Children's Passwords Using a Game-based Survey. In *Conference on Interaction Design and Children*, IDC '17, pages 617–622, Stanford, California, USA, June 2017. ACM.

[15] M. Lutaaya and S. Chiasson. Poster: Password Rehearsal Memory Games. In *USENIX Symposium on Usable Privacy and Security*, SOUPS '15, Ottawa, Ontario, Canada, July 2015. USENIX.

[16] N. Micallef and N. A. G. Arachchilage. A Gamified Approach to Improve Users' Memorability of Fall-back. In *The Who Are You?! Adventures in Authentication Workshop*, WAY '17, Santa Clara, California, USA, July 2017. USENIX.

[17] T. Seitz and H. Hussmann. PASDJO: Quantifying Password Strength Perceptions with an Online Game. In *Australian Conference on Human-Computer Interaction*, OzCHI '17, Brisbane, Australia, Nov. 2017. ACM.

[18] R. Shay, L. Bauer, N. Christin, L. F. Cranor, A. Forget, S. Komanduri, M. L. Mazurek, W. Melicher, S. M. Segreti, and B. Ur. A Spoonful of Sugar?: The Impact of Guidance and Feedback on Password-Creation Behavior. In *ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2903–2912, Seoul, Republic of Korea, Apr. 2015. ACM.

[19] B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. L. Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, N. Christin, and L. F. Cranor. How Does Your Password Measure Up? The Effect of Strength Meters on Password Creation. In *USENIX Security Symposium*, SSYM '12, pages 65–80, Bellevue, Washington, USA, Aug. 2012. USENIX.

[20] B. Ur, S. M. Segreti, L. Bauer, N. Christin, L. F. Cranor, S. Komanduri, D. Kurilova, M. L. Mazurek, W. Melicher, and R. Shay. Measuring Real-World Accuracies and Biases in Modeling Password Guessability. In *USENIX Security Symposium*, SSYM '15, pages 463–481, Washington, D.C., USA, Aug. 2015. USENIX.

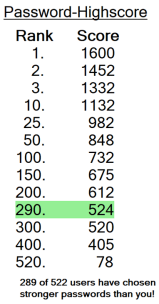
[21] A. Vance, D. Eargle, K. Ouimet, and D. Straub. Enhancing Password Security through Interactive Fear Appeals: A Web-based Field Experiment. In *Hawaii International Conference on System Sciences*, HICSS '13, pages 2988–2997, Wailea, Maui, Hawaii, USA, Jan. 2013. IEEE.

[22] S. P. Walz and S. Deterding. *The Gameful World: Approaches, Issues, Applications*. MIT Press, 2014.

[23] D. L. Wheeler. zxcvbn: Low-Budget Password Strength Estimation. In *USENIX Security Symposium*, SSYM '16, pages 157–173, Austin, Texas, USA, Aug. 2016. USENIX.

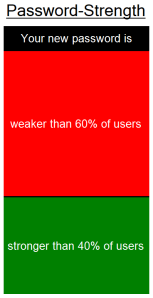
# APPENDIX

## A. TESTED CONDITIONS



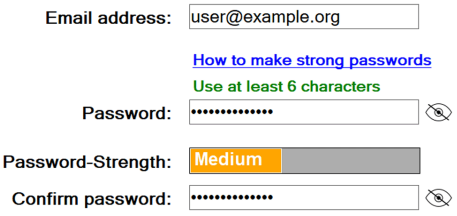

(a) 1: High Score Meter


(b) 2: Badges Meter

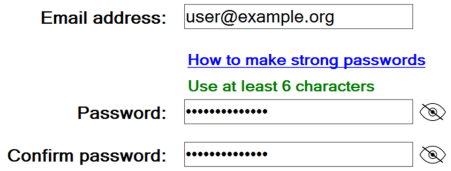



(c) 3: Peer-Pressure Meter

**Figure 1: The three non-bar conditions: First, a high score, prefilled with password strength scores derived from leaked real-world passwords. Second, a strength meter using nine different badges from a fantasy video game theme (based on designs by Krzysztof Dycha [7]). Showing three different categories, namely length, strength, and blacklist. All badges include a label that describes the purpose of the badge. Third, for comparability, a reimplementation of the peer-pressure meter from Egelman et al. [9].**




(a) 4: Bar Meter




(b) 5: No Meter

**Figure 2: The two control conditions: First, a traditional bar meter, changing its size and color from red (0-30 %) over yellow (40-60 %) to green (70-100 %) supported by a label displaying [Weak, Medium, Strong]. Second, an empty password creation form, showing no password strength meter.**

# B. QUESTIONAIRE

**Table 2: Detailed results of the questionnaire and demographic information. Questions Q2-Q5 were only shown to participants who saw a strength meter. We only considered participants that reported to not have participated before (cf. Q9).**

| | Questionnaire | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | High Score | | Badges | | Peer-Pressure | | Bar | | No Meter | | All | |
| | No. | % | No. | % | No. | % | No. | % | No. | % | No. | % |
| **Q1 – How many hours per week do you spend playing video games (console, PC, smartphone, etc.)?** | 62 | 100 % | 57 | 100 % | 65 | 100 % | 55 | 100 % | 63 | 100 % | 302 | 100 % |
| None | 12 | 19 % | 24 | 42 % | 12 | 18 % | 20 | 36 % | 14 | 22 % | 82 | 27 % |
| < 1 h | 12 | 19 % | 6 | 11 % | 12 | 18 % | 8 | 15 % | 15 | 24 % | 53 | 18 % |
| 1 - 3 h | 14 | 23 % | 7 | 12 % | 18 | 28 % | 8 | 15 % | 11 | 17 % | 58 | 19 % |
| 3 - 8 h | 14 | 23 % | 10 | 18 % | 12 | 18 % | 9 | 16 % | 16 | 25 % | 61 | 20 % |
| > 8 h | 10 | 16 % | 10 | 18 % | 11 | 17 % | 10 | 18 % | 7 | 11 % | 48 | 16 % |
| **Q2 – How much influence did the password strength meter have on your password choice?** | 62 | 100 % | 57 | 100 % | 65 | 100 % | 55 | 100 % | - | - | 239 | 100 % |
| No effect | 13 | 21 % | 18 | 32 % | 19 | 29 % | 16 | 29 % | - | - | 66 | 28 % |
| Minor effect | 12 | 19 % | 11 | 19 % | 14 | 22 % | 6 | 11 % | - | - | 43 | 18 % |
| Neutral | 12 | 19 % | 9 | 16 % | 6 | 9 % | 6 | 11 % | - | - | 33 | 14 % |
| Moderate effect | 18 | 29 % | 11 | 19 % | 16 | 25 % | 17 | 31 % | - | - | 62 | 26 % |
| Major effect | 7 | 11 % | 8 | 14 % | 9 | 14 % | 10 | 18 % | - | - | 34 | 14 % |
| Don't know | 0 | 0 % | 0 | 0 % | 1 | 2 % | 0 | 0 % | - | - | 1 | 0 % |
| **Q3 – I feel that the password strength meter shown uses an appropriate method to visualize password strength.** | 62 | 100 % | 57 | 100 % | 65 | 100 % | 55 | 100 % | - | - | 239 | 100 % |
| Strongly agree | 5 | 8 % | 3 | 5 % | 8 | 12 % | 9 | 16 % | - | - | 25 | 10 % |
| Agree | 34 | 55 % | 36 | 63 % | 40 | 62 % | 26 | 47 % | - | - | 136 | 57 % |
| Neither agree nor disagree | 14 | 23 % | 14 | 25 % | 10 | 15 % | 14 | 25 % | - | - | 52 | 22 % |
| Disagree | 3 | 5 % | 2 | 4 % | 6 | 9 % | 4 | 7 % | - | - | 15 | 6 % |
| Strongly disagree | 4 | 6 % | 1 | 2 % | 0 | 0 % | 2 | 4 % | - | - | 7 | 3 % |
| Don't know | 2 | 3 % | 1 | 2 % | 1 | 2 % | 0 | 0 % | - | - | 4 | 2 % |
| **Q4 – In your own words, please describe how the shown password strength meter reacts if a strong password is entered.** | 62 | 100 % | 57 | 100 % | 65 | 100 % | 55 | 100 % | - | - | 239 | 100 % |
| | free text | | free text | | free text | | free text | | - | - | free text | |
| **Q5 – How would you feel if your most visited website used this password strength meter?** | 42 | 100 % | 57 | 100 % | 65 | 100 % | 55 | 100 % | - | - | 239 | 100 % |
| Not at all satisfied | 3 | 7 % | 5 | 9 % | 8 | 12 % | 10 | 18 % | - | - | 26 | 11 % |
| Slightly satisfied | 12 | 29 % | 15 | 26 % | 12 | 18 % | 8 | 15 % | - | - | 47 | 20 % |
| Moderately satisfied | 1 | 2 % | 12 | 21 % | 18 | 28 % | 18 | 33 % | - | - | 69 | 29 % |
| Very satisfied | 22 | 52 % | 14 | 25 % | 17 | 26 % | 14 | 25 % | - | - | 67 | 28 % |
| Extremely satisfied | 1 | 2 % | 7 | 12 % | 6 | 9 % | 1 | 2 % | - | - | 15 | 6 % |
| Don't know | 3 | 7 % | 4 | 7 % | 4 | 6 % | 4 | 7 % | - | - | 15 | 6 % |

| | Demographics | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Q6 – How old are you?** | 62 | 100 % | 57 | 100 % | 65 | 100 % | 55 | 100 % | 63 | 100 % | 302 | 100 % |
| 18-20 | 28 | 45 % | 18 | 32 % | 26 | 40 % | 21 | 38 % | 29 | 46 % | 122 | 40 % |
| 21-25 | 28 | 45 % | 21 | 37 % | 24 | 37 % | 14 | 25 % | 19 | 30 % | 106 | 35 % |
| 26-30 | 5 | 8 % | 13 | 23 % | 11 | 17 % | 12 | 22 % | 14 | 22 % | 55 | 18 % |
| 31-40 | 0 | 0 % | 4 | 7 % | 4 | 6 % | 6 | 11 % | 0 | 0 % | 14 | 5 % |
| 41-50 | 1 | 2 % | 1 | 2 % | 0 | 0 % | 1 | 2 % | 1 | 2 % | 4 | 1 % |
| > 50 | 0 | 0 % | 0 | 0 % | 0 | 0 % | 1 | 2 % | 0 | 0 % | 1 | 1 % |
| **Q7 – With which gender do you identify?** | 62 | 100 % | 57 | 100 % | 65 | 100 % | 55 | 100 % | 63 | 100 % | 302 | 100 % |
| Female | 22 | 35 % | 20 | 35 % | 21 | 32 % | 23 | 42 % | 29 | 46 % | 115 | 38 % |
| Male | 38 | 61 % | 36 | 63 % | 41 | 63 % | 29 | 53 % | 32 | 51 % | 176 | 58 % |
| Other | 0 | 0 % | 0 | 0 % | 3 | 5 % | 0 | 0 % | 0 | 0 % | 3 | 1 % |
| Prefer not to answer | 2 | 3 % | 1 | 2 % | 0 | 0 % | 3 | 5 % | 2 | 3 % | 8 | 3 % |
| **Q8 – Are you majoring in, hold a degree in, or have held a job in any of the following fields: computer science; computer engineering; information technology; or a related field?** | 62 | 100 % | 57 | 100 % | 65 | 100 % | 55 | 100 % | 63 | 100 % | 302 | 100 % |
| Yes | 12 | 19 % | 8 | 14 % | 15 | 23 % | 13 | 24 % | 17 | 27 % | 65 | 22 % |
| No | 44 | 71 % | 46 | 81 % | 47 | 72 % | 37 | 67 % | 45 | 71 % | 219 | 73 % |
| Prefer not to answer | 6 | 10 % | 3 | 5 % | 3 | 5 % | 5 | 9 % | 1 | 2 % | 18 | 6 % |
| **Q9 – Have you already participated in this exact study or knew the true goal of it? (Be honest, you will get your compensation independent of what you answer)** | 67 | 100 % | 69 | 100 % | 74 | 100 % | 62 | 100 % | 70 | 100 % | 342 | 100 % |
| Yes | 5 | 7 % | 12 | 17 % | 9 | 12 % | 7 | 11 % | 7 | 10 % | 40 | 12 % |
| No | 62 | 93 % | 57 | 83 % | 65 | 88 % | 55 | 89 % | 63 | 90 % | 302 | 88 % |