An Evaluation of the Efficacy of Password Strength Meters

CPSC 525 Project Final Report – Group 10

Mark De Castro
University of Calgary
Calgary, Alberta
mdecastr@ucalgary.ca

Masih Sadat
University of Calgary
Calgary, Alberta
masih.sadat2@ucalgary.ca

Niroojen Thambimuthu
University of Calgary
Calgary, Alberta
niroojen.thambimuthu@ucalgary.ca

Cole Towstego
University of Calgary
Calgary, Alberta
ctowsteg@ucalgary.ca

ABSTRACT

Passwords remain the most common form of user authentication. As attackers use more sophisticated forms of attacks to crack passwords, it is imperative that users are able to select passwords which are strong and secure. The most common tool to assist with this is the strength meter, prevalent among numerous websites. However, the exact implementation of these meters can vary greatly, potentially confuse the user, and may not effectively evaluate a password's strength. We examine the password meters of various high-profile websites and conduct surveys to assess the impact of these meters on password strength. We then recommend a minimal list of requirements for an effective meter as per our results.

1 INTRODUCTION AND PROBLEM

We use passwords to protect the information we store on various web services, ranging from our social media accounts to the sensitive data found in our bank accounts. However, users of such websites may not always be aware of what a strong password may be, or may have no motivation to create one as long as they are able to immediately access these services. It then becomes the task of these websites to aid such users in formulating secure passwords which will help protect the information they store from possible threats or attacks. They commonly use password strength meters to do so. These are often the coloured bars that provide visual feedback to a user where in the range of "weak" to "strong" their proposed password lies. They are often accompanied by a set of guidelines of what the user's password should be comprised of.

A number of studies have been conducted in the past to assess the integrity and efficacy of these password meters. In 2014, de Carnavalet and Mannan conducted a large-scale analysis of the password meters of high-profile websites where they concluded that such meters are "highly inconsistent, fail to provide coherent feedback, and sometimes provide strength measurements that are blatantly misleading [1]." Wang and Wang in 2015 conducted a similar empirical analysis of the password policies of 50 leading web services, and found that they likewise provide highly inconsistent outcomes under identical testing, and that they "largely fail to withstand online guessing attacks [2]." Similarly, Ur *et al.* state that, while they found that these meters do affect user behaviour, "the resulting passwords were only marginally more resistant to password cracking attacks [3]."

Therefore, it is apparent that there is a serious problem among these password strength meters, in that they vary greatly in implementation, which can lead to incoherent or confusing feedback, and that they may not actually be correctly assessing a password's strength. These findings thus weaken the purpose of these password meters. This poses an alarming security risk, as this can allow individuals to have the wrong perception of what a strong password is. Consequently, with the ever growing number of attacks happening among websites today, individuals may then be more susceptible to password cracking and guessing attacks, and personal information leaks may become more prevalent. Thus, as noted by the National Institute of Standards and Technology (NIST), the use of effective password management "reduces the risk of compromise of password-based authentication systems [4]." As such, with a constantly growing online world, it is then relevant to analyze and reassess these password meters to see if there have been changes made in increasing their integrity and security, and if they are indeed the effective tools they ought to be.

2 APPROACH

In this project, we have repeated and extended upon parts of the empirical analyses and studies on the password strength meters of many popular websites, as found in the three cited papers above, which are by: (1) de Carnavalet and Mannan (2014), (2) Wang and Wang (2015), and (3) Ur *et al.* (2012).

To perform our evaluation, we have chosen a number of popular, high-profile websites to work with (as based on their rankings from *Alexa Internet*, which ranks web pages according to their web traffic (i.e., the time users spend on the site, number of daily page views, number of connections to other sites, etc.) over three-month periods). We primarily chose those with password strength meters and/or those that provide feedback to their users, as the evaluations will be based on the efficacy of these meters and feedback they give to a user. Finally, we also aimed to select a diverse list of websites, encompassing different areas ranging from social networks to online retailers, so that our study is representative enough of the large number of web services available today. With these conditions in mind, we have thus selected the following ten websites:

Table 1: Websites Chosen for Evaluation

Website	Alexa Ranking	Type
Apple	75	Various
Dropbox	89	File hosting service
eBay	44	Online retailer
Facebook	3	Social networking
FedEx	466	Courier/delivery service
Google	1	Web portal
Reddit	6	Discussion website
Twitch	32	Live video streaming
Twitter	13	Social networking
Wordpress	53	Blogging tool

Given this diverse list of services, we have conducted two tests to assess the password meters of our selected websites based on the work of de Carnavalet and Mannan and of Wang and Wang. These include: (1) an in-depth analysis of the characteristics of the password meters, and (2) a test to measure their resistance towards guessing attacks. Then, upon gathering insights based on these evaluations, we conducted surveys to assess the efficacy of two different password meters using two groups of people. We then gathered their responses afterwards with regard to their sentiments towards the use of such password meters.

3 EXPERIMENTS AND RESULTS

Our investigations begin with analyses based on the work of de Carnavalet and Mannan and those by Wang and Wang, which then conclude with work based on the study of Ur *et al.* These are detailed below:

3.1 Analysis of Password Meter Characteristics

For the first part of our evaluation, we analyzed the prominent characteristics of the password meters of our chosen websites, as well as the policies or guidelines they promulgate for password creation. This is based on the work done by de Carnavalet and Mannan in 2014, where they systematically characterized 22 password meters used by popular web services (e.g., Google, Microsoft) and leading password managers such as 1Password and LastPass. In their findings, they described these meters as "highly inconsistent" and "blatantly misleading," and have noted that many of them were "quite simplistic in nature" and "bear no indication of any serious efforts" in their design or implementation [1]. As such, we performed a similar assessment on our ten chosen meters to see how such popular websites have implemented their password meters, observing the similar characteristics sought out by the previous study, as well as any other attributes we deem pertinent for assessing the meters' efficacy. We looked for any peculiarities among them, which may lead us to finding any weaknesses or inconsistencies these meters may possess. We then compared our findings with those found in any similar services studied by de Carnavalet and Mannan. For our evaluation, we have observed the following characteristics:

- *Strength scale.* Does the meter use some form of visual scale/bar, or words such as "weak" or "strong" to rate a password's strength?
- *Length minimum and maximum.* What is the number minimum/maximum amount of characters allowed for a password, if any?
- *Character set requirements.* Are users required to use either numbers, uppercase or lowercase letters, or symbols in their passwords?
- *Monotonicity.* Does adding more characters increase a password's score?
- *Allows user information.* Can a personal information be used as a password (e.g., user's name, birthday, username they are using to sign up for the account)?
- *Allows spaces.* Do they allow the use of trailing spaces internally and externally?
- *Enforcement.* What is the minimum strength required for a password to be accepted?
- *Provides feedback.* Does any feedback popup to help the user as they create their password?

Table 2: Password Meter Characterization

Website	Strength scale	Length min/max	Charset required	Allows user info?	Allows spaces?	Enforcement	Feedback?
Apple	Weak, Moderate, Strong; Uses visual bars (from red, orange, green)	8 / 32	1+ Lower, 1+ upper, 1+ number	N	N / N	Moderate (or anything orange)	Y
Dropbox	Uses visual bars (0 to 4)	6 / 72	None	Y	Y / Y	None	N
eBay	Weak, Medium, Strong	6 / 64	1+ letter, 1+ number or symbol	Y	N / Y	Strong	Y
Facebook	Too short, Too weak, Strong; text changes colour from red to green	6 / –	None	Y	Y / Y	Strong	Y
FedEx	Uses visual bars (red, yellow, green)	8 / 35	1+ Lower, 1+ upper, 1+ number	N	N / N	Yellow	Y
Google	Too Short, Weak, Fair, Strong; Uses a visual bar that changes colour	8 / 100	None	N	N / Y	Fair	Y
Reddit	Weak, Fair, Good, Strong; Uses a meter that turns from red to green	6 / –	None	Y	Y / Y	Weak	Y
Twitch	Weak, Fair, Strong; Uses visual bars (1 to 5)	8 / –	None	Y	Y / Y	Weak	N
Twitter	Uses a green visual bar that fills up	6 / –	None	Y	Y / Y	If bar is halfway full	Y
Wordpress	Too short, Very Common, Too Easy to Guess, Accepted (Red to Green)	6 / –	None	N	Y / Y	Accepted	Y

Results. Table 2 summarizes our observations on the pertinent characteristics displayed by our ten password meters. Upon analysis, we have found the following:

- All the password meters vary greatly in terms of the strength scale that they use. Not all utilize some form of visual or coloured bar to suggest a password's strength. Also, their strength descriptions (i.e., "weak," "too weak," "strong") all differ from one another. Most of them have minimal explanation of how their meter works, or of how a strength score is assigned.
- Most employ a minimum of 6 characters, while half do not employ a maximum limit. Interestingly, both Apple and Google (both of which provide a wider range of services than the other sites) have a minimum of 8 characters instead. This is more desirable since, as per NIST's findings in their *Guide to Enterprise Password Management*, increasing a password's length (instead of its complexity, such as by adding various symbols) increases its security [4].
- Only three of the ten websites enforce the use of special characters. As per our previous observation, despite this low number, this is acceptable as adding uppercase letters, symbols or numbers do not necessarily increase a password's security.
- More than half of the meters allow the use of a user's personal information. This is alarming since, if an attacker gains any such information from a user (e.g., by phishing, shoulder surfing or social engineering), they will be able to guess their password more easily. This is even more disconcerting since, as per NIST's guide that we previously noted [4], as well as in their *Digital Identity Guidelines* [5], it has been found that users tend to use personal information when formulating passwords as this allows for better memorization. Thus, user information that is used for account creation should never be allowed to be used as, or within, one's password.
- Most of the websites accept passwords even if they have rated them as "weak" or "moderate." This is very dubious, since, for example, if a user enters in a

password deemed "weak" but it still gets accepted, they are then susceptible to having misconceptions of what a "weak" or "strong" password truly is. This can also lead to user frustration if, suppose, a user goes around using the same password on various sites; for example, Apple rates a password like "qwER43@!" as "strong," while others like Dropbox and Twitch rate it as "fair" or moderate.
- Only two of the ten websites do not provide a user any form of feedback as they enter in their desired password. This is excellent since (as per our analysis in section 3.3. below) users are less frustrated when the meter guides them through the password creation process through appropriate feedback. For example, most of these websites usually tell a user to avoid using easy to guess passwords (e.g., "123456" or "password"). Also noteworthy is FedEx, since they have even set up a separate page for a password creation guide, which provides techniques on how to create a stronger password.

Comparison of Results. Using the above findings, we can note that, as per the previous studies of de Carnavalet and Mannan's in 2014, in terms of diversity, none of the meters of the web services we have evaluated use a common meter. That is, they are, unfortunately, still highly inconsistent in that their implementations still vary greatly. In terms of the websites studied, both our studies have observed the meters found in Dropbox, FedEx, Twitter, eBay, Google and Apple (5 out of 10 our sites). One of the most notable changes between the observations is that of Twitter. In the 2014 study, it previously rated passwords from: 'invalid/too short,' 'obvious,' 'not secure enough,' 'could be more secure' 'okay,' to 'perfect.' Today, they instead use a green-coloured bar that fills up as a user's desired password is rated more strongly. There is also barely any feedback provided now, only telling a user to "enter a stronger password" if they deemed it weak. Likewise, FedEx also abandoned text ratings and instead now use three coloured bars. However, other requirements remained the same. Finally, eBay, Google and Apple, aside from a few minor tweaks (e.g., eBay increased the maximum character limit from 20 to 64,), nothing notable has been changed between our evaluation and that of de Carnavalet and Mannan of their password meters.

Thus, glaring inconsistencies and weaknesses are still present among the password meters of these popular web services. Only two of the ten websites (Apple and FedEx) are very stringent in terms of the requirements they have for password creation; however, they are still not flawless (i.e., they accept passwords that are deemed moderate). On the other hand, two of the ten seem to have placed no serious efforts in implementing their meters (Dropbox and Reddit) by posing security-threatening weaknesses, such as accepting passwords deemed "weak," combined with very relaxed requirements for character requirements, and even allowing the use of one's personal information. These disparate results are highly regrettable, as many users tend to utilize a number of these services simultaneously, and as noted above, user confusion (i.e., with regard to users being made aware of what a "strong" password is) and frustration can arise as these disparities and inconclusive strength ratings continue to exist.

3.2 Resistance to Guessing

One of the most common techniques that attackers use to infiltrate a user's online account is by *guessing*, which involves repeated attempts to authenticate falsely as the user by using common passwords and dictionary words. One form of guessing is through a *dictionary attack*, where the attacker guesses a password based on a list of probable passwords which people may have the tendency of using. Therefore, to counter such forms of attacks, NIST promulgates that organizations should make sure that users are disallowed from using "trivial passwords, . . . , simple keyboard patterns (e.g., "qwerty", "1234!@#$"), dates (e.g., "03011970"), dictionary words, and names of people and places" as their password during account creation [4]. It is therefore recommended that password meters and password creation policies utilize some form of blacklist of such common passwords to avoid the creation and use of vulnerable and ineffective passwords.

Thus, as an extension upon similar work performed by Wang and Wang in 2015, we have tested if our ten chosen websites's password meters, during account creation, resist very common passwords. For our tests, we want to find out if our ten meters will be able to prevent the use of any of the top 50 passwords from the "Worst Passwords of 2017," released by *SplashData*, a security firm that annually releases such lists of the most commonly hacked passwords as based on their examination of millions of passwords which have been leaked in data breaches [6]. Prior to testing, we assume that all ten web services should successfully reject most, if not, all, of the passwords found in this list since, by security standards, these are the most ineffective passwords that one can use to prevent their account from being compromised. We manually entered all 50 passwords on each of the ten websites' meters, and have consolidated the results in Table 3 below, registering their success rate in rejecting these weak passwords (i.e., number of rejected passwords divided by total number of passwords). A full table of the results can be found in Appendix A.

Table 3: Success Rate against 50 Worst Passwords of 2017

Website	Success (%)	Website	Success (%)
Apple	100	Google	100
Dropbox	12	Reddit	10
eBay	98	Twitch	68
Facebook	100	Twitter	86
FedEx	100	Wordpress	100

Results. To gain full insights on the success rate of our ten chosen meters, let us first analyze our list of 50 weak passwords. Indeed, this list can be considered as a blacklist of passwords which follows the aforementioned NIST guideline of passwords that should never be used. That is, the list indeed contains passwords that are: trivial (e.g., *blahblah*, *12341234*), simple keyboard patterns (e.g., *qwerty*, *qazwsx*), dictionary words (e.g., *password*, *football*, *monkey*), names of people (e.g., *robert*, *matthew*, *jordan*).

Now, upon analyzing our full list of results of which passwords were accepted or rejected, as well as Table 3 above, we have found that half of the password meters were able to successfully reject 50 of the worst/most common passwords of 2017. It is then refreshing to know that websites such as Apple and Google (which contain large amounts of very sensitive user information, such as personally identifiable information, emails, contacts, banking information, etc.) were all able to score a 100% success rate in disallowing these passwords. Likewise, for eBay, with the exception of one (i.e., *jordan23*, which does follow all its requirements), 98% of the passwords were rejected. However, Twitter and Twitch had lower success rates, having 86% and 68% respectively. For Twitter, those that it had accepted (7 out of 50) were very simple combinations of characters (e.g., *123456789*, *passw0rd*). Twitch had a lower success rate (accepting 16 out of 50) since, aside from accepting similar (weak) combinations of characters (e.g., *12341234*, *trustno1*), it also allowed a fair amount of random (yet common) phrases and words (e.g., *starwars*, *iloveyou*). However, most concerning are the results of Reddit and Dropbox. Reddit scored the lowest, rejecting only 5 of the 50 passwords. It is pertinent to note that these 5 rejections were solely based on the fact that they did not meet Reddit's requirement of at least 6 characters. It accepted *12345*, *admin*, *login*, *hello* and *1234* as legitimate passwords only because they were less than 6 characters. Otherwise, all other passwords were accepted, including very bad passwords like *password* and *123456*, which have been the top two worst passwords ever since *SplashData*'s first list dating back to 2011. While Reddit's low score is fairly reasonable, considering that the it stores fairly non-sensitive data (that is, they are mostly peoples' public discussions and conversations of on discussion boards or "subreddits," and that people who register an account in Reddit do so on anonymous terms), it should still have stricter policies since a breach of account information on Reddit could potentially be used to access a user's other accounts on other web services, as people have the tendency to reuse passwords for the sake of better memorization [5]. However, what is most alarming among all results is that Dropbox's score is one of the lowest, having rejected just one more than Reddit (i.e., *qwerty*). As Dropbox is a file hosting service, it holds large amounts of private data. Thus, Dropbox (being one of, if not the most prominent file sharing and storage services) having such an abhorrent success rate in rejecting common and very weak passwords is very alarming as it inspires an inferior model of password security. It should, in principle, have had the same results as other prominent companies like Google and Apple.

Therefore, with an overall success rate of 50% (i.e., 5 out of our 10 password meters unanimously rejected the passwords), our evaluation did not meet our assumption that all 10 should have at least scored high success rates. That is, regardless of type of web service, a password meter should, in principle, always reject passwords which can be easily guessed by attackers. Thus, with these results, we can state that some of these popular websites' password meters are largely unsuccessful in providing some of the most basic forms of security measures with regard to account protection.

Comparison of Results. In 2015, Wang and Wang conducted a similar evaluation, where they tested 16 passwords (such as *123456*, *password*, *iloveyou*, and other weak combinations of letters and symbols) on 50 websites, for a total of 800 testing instances [2]. They found that only 259 instances were rejected, (32% success rate of password rejection). Thus, they conclude that the meters they tested "largely fail to serve their purposes" of providing security. Inspired by their methods, we have extended upon their study, and having evaluated 50 similar weak passwords against various popular websites, our results are very similar to theirs. Both our tests scored low overall success rates (our 50% versus their 32%) of password meters rejecting common passwords. Like our study, they have also evaluated Apple, Google (via Gmail), Twitter and Facebook, and have also used three of the same passwords we used on our worst password list, which are: *123456* (#1), *password* (#2) and *iloveyou* (#10). Comparing results, both studies have shown that all four prominent websites reject these three passwords, as well as other very common and weak passwords.

3.3 Survey on the Presence of Password Meters

Having gained insights on what weaknesses and inconsistencies are present among the password meters of popular websites (as based on our two previous analyses), we now extend upon the work of Ur *et al.*, who, in 2012, have conducted a 2,931-subject study of password creation using 14 password meters to find out if such meters do affect how users create their passwords. In this study, they asked participants to create a password on one of the 14 meters, then asked them to complete a survey about how they handled their password using those meters [3]. As such, we have conducted a similar study to determine whether or not the presence of a password meter would have a measurable impact on the strength of the passwords created by our users. To accomplish this, we created two different variants of our survey: one with a password meter, and one without. Each variant also had its own set of questions which served to help us understand the sentiments of the users with regard to password meters and password security. We then used Amazon Mechanical Turk to gather 100 responses for each version of our survey.

Tooling. As our survey was web-based, we made use of the Bootstrap framework to piece together a relatively modern looking webpage which would seem familiar to what a user might expect when a user is creating an account or changing their password. For our password meter, we used Dan Wheeler's *zxcvbn* tool, an open-source password strength estimator, which analyses a password based on various patterns and dictionaries. It then calculates an entropy and crack time for the password, then, based on these, gives it a strength score from 0 to 4 (with 4 as the highest) [7]. We used this tool as it was easy to implement (natively developed in JavaScript), and is generally considered as one of the best implementations of password meters, as it is capable of considering many factors such as common passwords, patterns, character substitutions, and personal information into its calculations of password strength. de Carnavalet and Mannan also note that *zxcvbn* yielded "more accurate strength evaluations" [1] than the other meters they studied. Once the survey was completed by the user, a random survey code would be generated by the server and returned to the user, to verify the completion of their work. Survey results were captured as an array of flat, one-dimensional JSON objects and stored on the server. Once the results were gathered, the results were processed into a CSV format through the use of a Python script, and then exported to Microsoft Excel for examination.

Survey Design and Rationale. When first accessing the survey, the user is prompted to suggest a password. The survey is designed to mimic a real-world scenario that a user might expect to see when they create an account. There are two masked text fields in which the user must enter their desired password. This approach is typically used to prevent the user from choosing a password which they did not intend by accident, but was also advantageous experimentally as it forces the user to select something which they can actually remember (i.e., at least long enough to type twice). This provides us with a better representation of realistic passwords, as the user cannot simply enter a long string of characters at random. Additionally, we imposed two other common restrictions: Users were not allowed to use common passwords (i.e. passwords with a *zxcvbn* score of 0), and passwords had to be a minimum of 8 characters long. We did not impose any requirements in regard to the use of special characters, uppercase characters or lowercase characters, nor did we forbid their use. Once the user had created a password that satisfied the requirements, they were able to move on to the next page of the survey in which they were asked to answer 5 multiple-choice questions. Two variants of our survey were created, both following this general format but differing on the following points:

- **Survey A:** A password meter is displayed while the user creates their password. The meter has five possible scores enumerated as {Very Weak, Weak, OK, Strong, Very Strong}. The meter also displayed a colored progress bar which correlates to the above scores as {(Red, 20%), (Red, 40%), (Orange, 60%), (Strong, 80%), (Very Strong, 100%)}. As each of the password requirements was met, a green tick would also appear next to the requirement to indicate that it has been satisfied. Once all requirements were satisfied, the user may proceed to the feedback page. Here, users would respond to each statement with one of {Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree} which were enumerated as scores from 1 to 5 (inclusively and respectively). The questions asked in this version were:
  1. The password meter was helpful in creating a strong password.
  2. Password meters are difficult or annoying to use.
  3. I understand how my password was rated by the meter.
  4. I would have made the same password without the presence of a password meter.
  5. I feel the meter gave my password an appropriate score.

- **Survey B:** Unlike survey A, a password meter was not displayed while the user is asked to create their password. The password requirements were identical to survey A, but no indicator was displayed to show that the conditions had been fulfilled. The user could only proceed to the multiple-choice phase once all requirements had been satisfied. Following the same scheme as survey A, the users were asked the following questions:
  1. The presence of a password meter would have helped in the creation of my password.
  2. Creating a password without any feedback was frustrating and/or annoying.
  3. I understand why my password was accepted (or rejected).
  4. I would have made the same password even if a meter were present.
  5. The password I submitted was "strong enough."

For screenshots of the two surveys, refer to Appendix B.

Results. Table 4 below summarizes the average calculated entropy and calculated score given by *zxcvbn* to each of the 100 results in our two surveys:

Table 4: Entropies and Scores from *zxcvbn*

	Survey A	Survey B
Average Entropy	38.41	38.09
Average Score	2.82	2.43

First, we explain how *zxcvbn* calculates these values briefly. As per its specifications, a password's entropy is calculated based on numerous patterns and dictionary matches found within a password, which *zxcvbn* checks for. Then, it uses this entropy to calculate a probable crack time of the password. This crack time then corresponds to scores from 0 to 4, where a score of 1 (the lowest possible score that we accepted) means the password's crack time is between 10^2 and 10^4 seconds, while a score of 4 (the highest) corresponds to a crack time greater than 10^8 seconds (about 3 years) [7]. We accepted even low scoring passwords since the point of these surveys are not to be stringent, but rather, to know how the presence of a meter and feedback affects the participants' password creation. Realistically, we would only accept at least a score of 3 or more to ensure proper password security. Now, in Table 4, while only marginally better, both the average entropy and average score given from *zxcvbn*'s calculations for survey A (i.e., those who used a password meter and had feedback) was higher than B (i.e., those who did not have a meter nor any feedback). However, if we look closer at each of the scores, 38% of the passwords in A scored a 4, while only 28% scored a 4 in B. On the other hand, 32% of the passwords in B scored a 1, while only 17% scored a 1 in A. Also, A scored more 3's than B, while B scored more 2's than A. Therefore, survey A's passwords generally scored better than those in B. As such, based on this empirical data produced by our surveys, having a password meter and having feedback present during password creation indeed leads to a stronger password (as per *zxcvbn*). For a complete list of created passwords and their scores, see Appendix C.

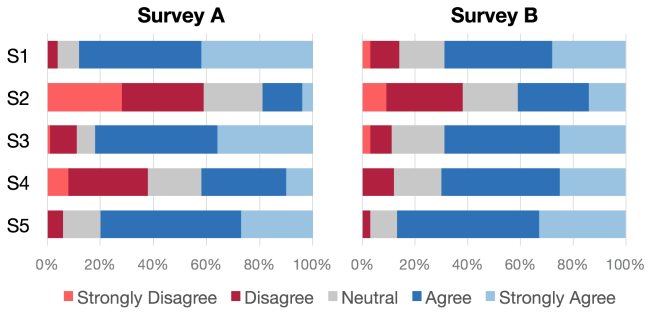

Figure 1. These charts depict participants' agreement or disagreement with each of the surveys' 5 statements (S1 to S5). Each color represents the proportion of participants in that condition who expressed a particular level of agreement or disagreement with the statement.

We now take a look at the sentiments of our participants with regard to how they felt during password creation by comparing each pair of questions from both surveys (e.g., statement 1 from both A and B). We computed for how frequent the participants agreed or disagreed toward either statement. Note that we combined 'Dis/agree' and 'Strongly Dis/agree' responses below for brevity's sake, while outliers are mentioned, if necessary. Using Figure 1 as a guide, the results are as follows:

- For statement 1, in survey A, 88% of the participants generally agreed that the meter was helpful for password creation. For B, 69% felt that they could have benefited from a meter.
- For statement 2, 59% in A generally disagreed that password meters are annoying. On the other hand, in B, 41% agreed that no meter nor feedback was frustrating, while 38% felt it was not.
- For 3, 82% and 69% in A and B respectively agreed that they understood how their password was rated by our meter (and non-meter).
- For 4, only 42% in A felt that they would have made the same password without the meter, while 70% in B felt they would have made the same password even with a meter.

- For 5, 80% and 87% felt in A and B respectively that their password was appropriately or strongly scored.

Based on the combined sentiments from statements 1, 2 and 4, all results point to the fact that participants felt that password meters (and the accompanying feedback) were helpful when they created their passwords. However, while those in A felt that the presence of a meter was not annoying, we obtained mixed results for B with regard to the absence of a meter being annoying, where about 40% both agreed and disagreed, averaging a Neutral response. As such, going by A's sentiments, we still believe having a meter would indeed lead to less user frustration or confusion. Then, for statements 3 and 5 for both surveys, a majority of participants in both A and B felt they understood how their password was rated and that it was appropriated rated by our meter (and accepted correctly by the one without a meter). Thus, since we have based our criteria (in A and B) and the look of our meter (in A) on the findings we gathered from our first two analyses in sections 3.1. and 3.2., we believe that the meter we used, as well as the requirements we set out, can be used as a good model for what an effective password can be.

Therefore, as per the results of the above evaluations (i.e., the scores given by *zxcvbn* and the general sentiments of our participants), we believe that the presence of password meters (and their accompanying guidelines) do indeed impact the password creation practices of users, and that they do lead to passwords which are generally stronger that those created in an environment without one.

Comparison of Results. In the previous findings of Ur *et al.*, they have stated that, when they used stringent-scoring meters (i.e., only accepted stronger passwords), "the resulting passwords were only marginally more resistant to password cracking attacks [3]." We have likewise found similar results, as the average scores given by *zxcvbn* for our surveys' passwords were between 2 and 3, and we would have gained higher scores if we only allowed scores of 3 or higher. They have also found that when they used meters that were "not overly onerous," rigorous meters indeed provided better security. This reflects our own findings that, with the use of a straightforward password meter, passwords tend to be stronger and more secure.

4 CONCLUSION

In this project, we have first conducted two analyses to assess the efficacy of the password meters deployed by popular web services. We have found that, like in past studies, even today, many of these meters are still highly inconsistent and possess various weakness prone to attacks. Then, using these results as a guide, we conducted surveys to assess if the presence of a meter impacts user sentiments and the strength of the passwords they create, and indeed, like our referenced study, they do indeed affect these during password creation. As such, based on all these findings, as well as with the evaluation standards of NIST that we have considered, we recommend that, for a password meter and its accompanying guidelines to be deemed effective, it should, at the minimum, have the following characteristics: (1) require a minimum of 8 characters (since the jump in the efficacy of brute force and guessing attacks from 6 to 8 characters is quite noticeable); (2) disallow the use of common and leaked passwords (and those found in publicly available blacklists); (3) disallow the use of personal information; and (4) use a straightforward visual meter. Therefore, we expect that the findings we have presented in this paper will aid in further improving the integrity and security of the many existing password meters today, and thus, allow them to be the effective tools that they are required to be.

REFERENCES

[1] Xavier de Carné de Carnavalet and Mohammad Mannan. Large-scale evaluation of high-impact password strength meters. *ACM Transactions on Information and System Security*, 18(1):1–32, June 2015.

[2] Ding Wang and Ping Wang. The emperor's new password creation policies: an evaluation of leading web services and the effect of role in resisting against online guessing. In *Proceedings of the 20th European Symposium on Research in Computer Security (ESORICS'2015)*, pages 456–477, Vienna, Austria, September 2015.

[3] Blase Ur, Patrick Gage Kelley, Saranga Komanduri, Joel Lee, Michael Maass, Michelle L. Mazurek, Timothy Passaro, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. How does your password measure up? The effect of strength meters on password creation. In *Proceedings of the 21st USENIX Security Symposium (USENIX Security'12)*, pages 65–80, Bellevue, Washington, USA, August 2012.

[4] Karen Scarfone and Murugiah Souppaya. NIST Special Publication 800-118: Guide to enterprise password management. Technical report, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, April 2009.

[5] Paul A. Grassi, Michael E. Garcia, and James L. Fenton. NIST Special Publication 800-63-3: Digital identity guidelines. Technical report, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, June 2017.

[6] TeamsID. 100 worst passwords of 2017. https://www.teamsid.com/worst-passwords-2017-full-list/. Accessed March 1, 2018.

[7] Daniel Lowe Wheeler. zxcvbn: low-budget password strength estimation. In *Proceedings of the 25th USENIX Security Symposium (USENIX Security'16)*, pages 157–173, Austin, Texas, USA, August 2016.

APPENDICES

A Results for Guessing Resistance

The table below provides an overview of data we have gathered with regard to whether or not our ten chosen websites' password meters reject the top 50 of the "Worst Passwords of 2017." Passwords are arranged by rank. (*Symbol:* 1 = Apple, 2 = Dropbox, 3 = eBay, 4 = Facebook, 5 = FedEx, 6 = Google, 7 = Reddit, 8 = Twitch, 9 = Twitter, 10 = Wordpress; Y = the password was accepted, N = the password was rejected).

Table 5: Do these password meters accept/reject top 50 worst passwords of 2017?

Password	1	2	3	4	5	6	7	8	9	10
123456	N	Y	N	N	N	N	Y	N	N	N
password	N	Y	N	N	N	N	Y	Y	N	N
12345678	N	Y	N	N	N	N	Y	Y	N	N
qwerty	N	N	N	N	N	N	Y	N	N	N
12345	N	N	N	N	N	N	N	N	N	N
123456789	N	Y	N	N	N	N	Y	Y	Y	N
letmein	N	Y	N	N	N	N	Y	N	N	N
1234567	N	Y	N	N	N	N	Y	N	N	N
football	N	Y	N	N	N	N	Y	Y	N	N
iloveyou	N	Y	N	N	N	N	Y	Y	N	N
admin	N	N	N	N	N	N	N	N	N	N
welcome	N	Y	N	N	N	N	Y	N	N	N
monkey	N	Y	N	N	N	N	Y	N	N	N
login	N	N	N	N	N	N	N	N	N	N
abc123	N	Y	N	N	N	N	Y	N	N	N
starwars	N	Y	N	N	N	N	Y	Y	N	N
123123	N	Y	N	N	N	N	Y	N	N	N
dragon	N	Y	N	N	N	N	Y	N	N	N
passw0rd	N	Y	N	N	N	N	Y	Y	Y	N
master	N	Y	N	N	N	N	Y	N	N	N
hello	N	N	N	N	N	N	N	N	N	N
freedom	N	Y	N	N	N	N	Y	N	N	N
whatever	N	Y	N	N	N	N	Y	Y	N	N
qazwsx	N	Y	N	N	N	N	Y	N	N	N
trustno1	N	Y	N	N	N	N	Y	Y	N	N
654321	N	Y	N	N	N	N	Y	N	N	N
jordan23	N	Y	Y	N	N	N	Y	Y	Y	N
harley	N	Y	N	N	N	N	Y	N	N	N
password1	N	Y	N	N	N	N	Y	Y	N	N
1234	N	N	N	N	N	N	N	N	N	N
robert	N	Y	N	N	N	N	Y	N	N	N
matthew	N	Y	N	N	N	N	Y	N	N	N
jordan	N	Y	N	N	N	N	Y	N	N	N
asshole	N	Y	N	N	N	N	Y	N	Y	N
daniel	N	Y	N	N	N	N	Y	N	N	N
andrew	N	Y	N	N	N	N	Y	N	N	N
lakers	N	Y	N	N	N	N	Y	N	N	N
andrea	N	Y	N	N	N	N	Y	N	N	N
buster	N	Y	N	N	N	N	Y	N	N	N
joshua	N	Y	N	N	N	N	Y	N	N	N
1qaz2wsx	N	Y	N	N	N	N	Y	Y	Y	N
12341234	N	Y	N	N	N	N	Y	Y	Y	N
ferrari	N	Y	N	N	N	N	Y	N	N	N
cheese	N	Y	N	N	N	N	Y	N	N	N
computer	N	Y	N	N	N	N	Y	N	N	N
corvette	N	Y	N	N	N	N	Y	Y	N	N
blahblah	N	Y	N	N	N	N	Y	Y	Y	Y
george	N	Y	N	N	N	N	Y	N	N	N
mercedes	N	Y	N	N	N	N	Y	Y	N	N
121212	N	Y	N	N	N	N	Y	N	N	N

B Survey Screenshots

The screenshots below are of the password meters and guidelines we used for our two surveys. The first is the survey where a meter was used and feedback was given, while the second has no meter and no feedback was given.

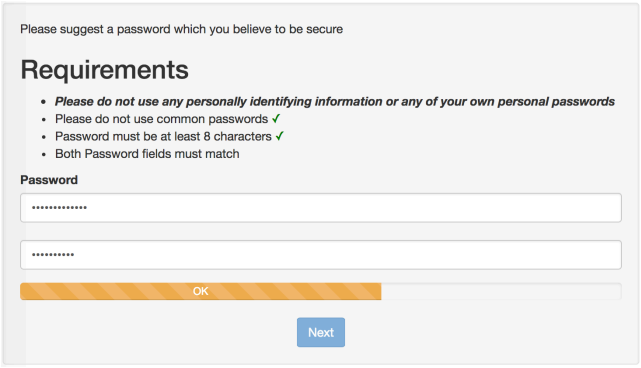

Figure 2. Survey with meter and feedback.

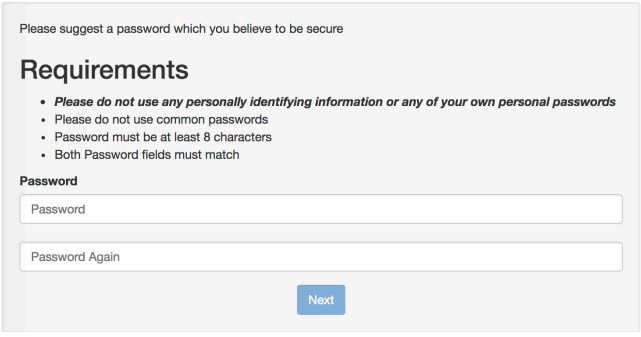

Figure 3. Survey with no meter and no feedback.

C Survey Results

The results in the tables below provide a look at subsets of the passwords created by 50 of the 200 participants in our two surveys, along with their entropies and scores provided by *zxcvbn*.

Table 6: Subset of Survey A Results

Password	Entropy	Score
PinkFrog135%L	45.032	4
Seetharaja	32.529	2
green1711CITY!&!!;'	49.114	4
Ekit4mb4l4	38.867	3
SHAMMu@420s	43.836	4
obiwan1220	22.872	1
EgbDf5454!	38.881	3
8489449971	33.219	2
@857bJ*Sl9	52.590	4
alphatangoboxer	31.974	2
Harshad@130	29.976	2
arunlovesindhu123	44.691	4
RIPO*143JAS	47.426	4
ass198230	24.689	1
$87?QAg7	49.311	4
@CasaBlanca12	31.303	2
theLO$94	38.433	3
14cme4914cme49	46.901	4
germankitties73	36.352	3
gocolts559	28.347	2
Straw2lover	27.898	2
96558495	26.575	1
22011990gaurav	37.900	3
RaM#J@i=861	43.835	4
PK1099life!	35.536	3

Table 7: Subset of Survey B Results

Password	Entropy	Score
summa2018	20.936	1
exf-sY2-Vbq-kEd	94.788	4
AllD@yIDr3am!$	33.534	2
KIRAN1000	24.947	1
monkeyhorsebattery	24.442	1
@Turuleto87%	46.203	4
dinkolfy@123	40.172	3
leethiyal	26.087	1
W9me62hk221Dar	66.901	4
81229037	26.575	1
5050505	21.976	1
Catcher85!	28.357	2
*20Spring18!	40.741	3
Rapid3l3v3n	26.150	1
ghserdre	30.143	2
kumarkumar12345	26.830	1
Dothedew1977	26.296	1
Loveme-2	22.747	1
Llemon1#1$	30.467	2
Marsattacks1	26.691	1
K8257Drive	30.197	2
X6jlol165&!	50.484	4
9600969442	33.219	2
c1i2n3d4y5	37.159	3
Karthi6009	35.419	3