Does My Password Go up to Eleven? The Impact of Password Meters on Password Selection

Serge Egelman[1], **Andreas Sotirakopoulos**[2],
Ildar Muslukhov[2], **Konstantin Beznosov**[2], **and Cormac Herley**[3]

[1]University of California, Berkeley
Berkeley, California
egelman@cs.berkeley.edu

[2]University of British Columbia
Vancouver, British Columbia
andreass,ildarm,beznosov@ece.ubc.ca

[3]Microsoft Research
Redmond, Washington
cormac@microsoft.com

ABSTRACT

Password meters tell users whether their passwords are "weak" or "strong." We performed a laboratory experiment to examine whether these meters influenced users' password selections when they were forced to change their real passwords, and when they were not told that their passwords were the subject of a study. We observed that the presence of meters yielded significantly stronger passwords. We performed a followup field experiment to test a different scenario: creating a password for an unimportant account. In this scenario, we found that the meters made no observable difference: participants simply reused weak passwords that they used to protect similar low-risk accounts. We conclude that meters result in stronger passwords when users are forced to change existing passwords on "important" accounts and that individual meter design decisions likely have a marginal impact.

Author Keywords
Security; Passwords; User Study

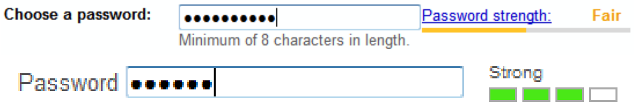
ACM Classification Keywords
D.4.6 Security and Protection: Authentication; H.1.2 User/Machine Systems: Human factors

General Terms
Security; Human Factors; Passwords

INTRODUCTION

> *If we need that extra push over the cliff, you know what we do?...Eleven. Exactly. One louder.*
> –Nigel Tufnel, *This Is Spinal Tap*

Password strength meters often appear when users create new accounts. In fact, of Alexa's top 20 websites [3], fifteen (75%) present users with meters during either password creation or changes. Such meters are updated in realtime to show where on the spectrum between "weak" and "strong" the proposed password lies (Figure 1). The implicit premise is that strong passwords are always desirable and that users who choose weak passwords do so because they are unaware that their passwords are weak; when made aware of weak passwords through a meter's feedback, there is an expectation that the user will choose a stronger password.



Minimum of 8 characters in length.

Figure 1. Password meters from Gmail and Yahoo!

Despite their ubiquity, we are unaware of prior research examining the effectiveness of password meters in situ. Ur *et al.* previously examined a variety of meter designs when it was known to users that passwords were the subject of the experiment [22]. As we will show in this work, their results are not reliable predictors of meter effectiveness because they did not account for the varying contexts in which meters are shown (nor did they study users' actual passwords). In this paper we performed two experiments, one in the laboratory and one in the field. We explored two different use cases: passwords used to protect sensitive accounts and passwords used to protect unimportant accounts. Across both use cases, we tested two types of meters: the traditional "weak" versus "strong" meter, as well as a new type of password meter that we developed to show password strength relative to other users on the system. Our contributions are as follows:

- We measured the extent to which password strength meters influenced users' password choices when they used their real passwords and were not told that passwords were the subject of the study.
- We show that password strength meters' influence on user behavior is heavily dependent on the context in which the password is used.
- We show that meters do not observably increase memorability problems, and postulate that such problems are more likely to be attributed to expiration policies.
- We show that the mere presence of a meter has greater impact than individual meter design decisions.

CHI 2013, April 27–May 2, 2013, Paris, France.
Copyright 2013 ACM 978-1-4503-1899-0/13/04...$15.00.

BACKGROUND

Despite attempts by many large sites to influence users with password strength meters, there has been very little study of them in the literature. The extent to which they influence users' password selections is largely unknown. We present an overview of prior research on password strength, password usability, and the use of soft paternalism to nudge users into making better decisions.

Password Strength

That user-chosen passwords fall into predictable patterns has been well documented. Morris and Thompson found that a large fraction of passwords on a Unix system were easily guessable [17]. Three decades later, Florêncio and Herley found that web users gravitate toward the weakest passwords allowed [11]. Several recent leaks of large password datasets have revealed that certain popular choices, such as "123456," are exceedingly common [24].

While much effort has been devoted to encouraging users to choose strong passwords, the concept of password strength remains surprisingly difficult to define. The natural measures, such as Shannon entropy or guessing entropy require knowing the probability distribution of passwords. Early efforts to quantify password strength resembled the measures of cryptographic strength: a password of length N, drawn from an alphabet of size C, would then have strength $N \log_2 C$ bits. NIST guidelines give a variation of this approach [23], where strength is a function only of length and character composition.

Weir *et al.* showed that neither of these measures offers a good guide to the resistance of a guessing attack [24], a finding corroborated by Shay *et al.* [19]. Password-cracking tools, such as John The Ripper [1], make heavy use of word-lists and achieve success far in excess of what the NIST entropy predicts [8]. Some passwords that appear strong under the early entropy measures fall relatively quickly to cracking tools. Probabilistic context-free grammars are likely to surpass even the best word-list based results [25]. While the concept of strength may be ill-defined, it appears clear that an ideal strength of a password would be an increasing function of the difficulty it presents to modern cracking tools.

Schechter *et al.* suggested that popularity of a password is the main predictor of weakness and suggested a data structure to limit the number of users who can use a password at a given site [18]. Bonneau proposed a novel measure and showed that it is a good predictor of attacker success over a corpus of 70 million Yahoo! passwords [5]. Yet none of these metrics are well-suited to strength meters because they either cannot be computed in realtime or they require web browsers to download unfeasibly large tables of probabilities. Castelluccia *et al.* suggested a new strength measure that could easily be incorporated into password meters [6]. They used an N-gram Markov model to predict characters and built an adaptive strength-meter. However, they did not validate their meter with real users, so the extent to which their new meter may influence user behavior is unknown.

While flawed as a metric for absolute password strength, zero-order entropy is a reasonable metric for examining the effectiveness of password meters because it is the metric on which many meters currently base their feedback [20]. Thus, we study password meters as they currently are, rather than as one might want them to be.

Password Usability

Many sites try to prevent weak password choices by enforcing password composition policies, which can also make passwords harder to remember [2]. While the usability burden is high, numerous attempts to replace passwords have accomplished little. In fact, Herley and van Oorschot argue that despite their shortcomings, passwords are unlikely to be supplanted soon [13].

The number of passwords that users must manage has increased, and thus usability has decreased. Florêncio and Herley found that the average user has 25 password-protected accounts [11]. To cope with this burden, most users reuse passwords between accounts.

Surprisingly, there has been little systematic analysis of how strength can be achieved with minimal usability impact. Yan *et al.* reported that mnemonic passwords are as good as random passwords in resisting brute-force attacks [14]. However, Kuo *et al.* found that with a properly chosen dictionary, the brute-forcing success rate increases dramatically [16].

Shay *et al.* studied users during the transition from a relatively relaxed to a very stringent policy [19]. Their participants found the transition very annoying, but perceived that security had improved. This correlates well with the anecdotal evidence that users find password composition policies particularly frustrating [7].

Komanduri *et al.* examined the impact that composition policies had on password strength and memorability [15]. They observed that longer passwords with no other requirements were significantly less onerous to users and resulted in stronger passwords, as compared to shorter requirements that mandated certain character classes (e.g., symbols, numbers, etc.).

Nudges

Thaler and Sunstein suggest that subtle encouragements, or nudges, can be effective at improving outcomes [21]. They posit that this is true for many economic and health problems, where mandates are difficult or undesirable, but poor user-choice can lead to bad effects. Passwords certainly provide an example where broadcasting suggestions on choosing strong passwords has not been successful. Indeed our work is in part inspired by the desire to determine whether a better usability-security tradeoff can be achieved by delivering a nudge in the form of password meter information about how a user-chosen password compares to those of peers.

Some of these techniques have started to be applied to solving computer security problems. For instance, Egelman *et al.* showed that framing can have a significant impact on user tolerance of security delays [10]. Besmer *et al.* showed that when framing access control decisions in terms of how their friends acted, users make significantly different decisions about what information to share [4].

Forget *et al.* investigated improving user-chosen passwords by providing optional system-generated modifications [12]. They found that users' initial choices were often weak, but they accepted the modifications, which significantly improved the zero-order entropy. Since the password is still derived from the original choice, they plausibly claim that the usability reduction is smaller than would be achieved by other approaches.

Most recently, Ur *et al.* examined the extent to which password meters influence users' password selections [22]. They examined 14 different meter designs and concluded that meters, regardless of specific design choices, resulted in significantly longer passwords over the control condition. Because participants did not use their actual passwords and understood that passwords were the subject of the experiment, their results represent a theoretical upper bound; they studied meter efficacy, whereas we study meter effectiveness. Thus, we are not aware that anyone has performed an ecologically valid study of password strength meters.

LABORATORY EXPERIMENT

We performed a between-subjects laboratory experiment in the fall of 2011 to test the following hypotheses:

H_0: Password are not stronger when meters are present.

H_1: Passwords are stronger when users see relative strength meters compared to no meters.

H_2: Passwords are stronger when users see relative strength meters compared to "traditional" meters.

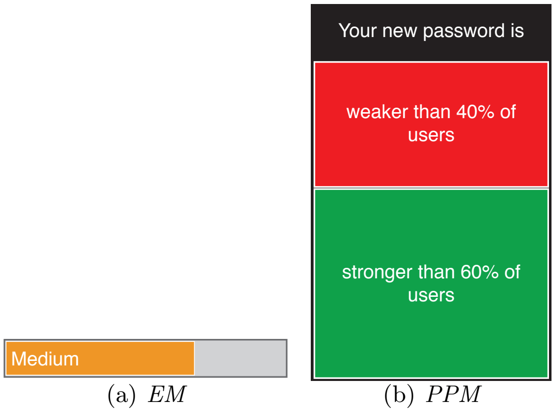
Our experiment involved 47 participants who changed passwords protecting important accounts. We examined how two different meters influenced password selection and memorability. In this section we describe our methodology, present our results, and discuss some open questions that this experiment was unable to answer.

Methodology

In this section we describe our experimental conditions, experimental protocol, and participants.

Conditions

The goal of our first experiment was to evaluate two types of password strength meters (Figure 2). One experimental condition featured a traditional meter that presented feedback in terms of whether the password was "weak," "medium," or "strong." We called this the "existing motivator" condition (EM). Our second experimental condition framed the password in terms of social pressure by presenting strength relative to all of the other users on the system. We called this the "peer-pressure motivator" condition (PPM). We randomly assigned participants to one of three between-subjects conditions: *EM*, *PPM*, or a *control* condition in which participants saw no password strength meter.



(a) *EM* (b) *PPM*

Figure 2. The password meters in the "existing motivator" (a) and "peer-pressure motivator" (b) conditions.

Our password meters used zero-order entropy as a metric for strength, calculated by the equation, $N \log_2 C$, where C is the character set size (e.g., 36 if it consists of numbers and lower-case letters), and N is equal to the length of the password. While this metric suffers from several limitations, the most serious of which is that character frequency is not considered, it is the metric upon which existing password strength meters are currently based [20]. Our interest was in examining whether users would choose longer passwords with more diverse character sets when presented with meters encouraging them to do so. Therefore, zero-order entropy was entirely appropriate for quantifying *relative* differences between conditions. We tested whether the meters yielded *stronger* passwords (i.e., longer with more character sets), not whether individual passwords were considered *strong* by themselves, which is why this metric fulfilled our needs.

Because we did not know the passwords of every user on the system we examined, we needed a way of calibrating the meters. We used the RockYou dataset by removing all passwords that did not meet our system's enforced minimum requirements—eight characters including one letter and one digit—and then we examined the median zero-order entropy of the remaining passwords. This median was then used to represent the "medium" level in the *EM* condition and the 50^{th} percentile in the *PPM* condition. In a pilot experiment ($n = 51$) we observed that almost all participants' initial passwords were well above this median, which meant that they had no reason to change their passwords based on the meters' feedback (i.e., the meters indicated that their initial passwords were strong). Because of this, we artificially inflated our thresholds to yield the intervals shown in Table 1.

Protocol

One concern when conducting security usability studies is that participants may not behave as they normally would if they are aware of the study's true purpose. Specifically, security is a secondary task; most users do not sit down at the computer to "do security." Thus, to maximize external validity, we ensured that partici-

Bit Strength (x)	PPM	EM
x<=53.41	0%	Weak
53.41<x<=56.53	30%	Weak
56.53<x<=59.83	40%	Medium
59.83<x<=64.26	50%	Medium
64.26<x<=71.09	60%	Medium
71.09<x<=77.21	70%	Strong
77.21<x<=82.27	80%	Strong
82.27<x<=83.30	90%	Strong
83.30<x	100%	Strong

Table 1. Password strength intervals used to provide feedback during the laboratory experiment.

pants created passwords for accounts that they actually cared to protect. To satisfy these constraints, we limited participation to affiliates of the University of British Columbia, which maintains a single sign-on (SSO) system for use by students, faculty, and staff. SSO accounts are used to perform tasks such as checking email, checking out books, viewing grades, and various other sensitive activities. SSO accounts are also used to access a campus portal. We told participants that we were studying the usability of this portal.

Participants logged in to the portal with their real passwords. We routed traffic through a proxy server in order for us to collect data. Upon successful login, the proxy server injected a dialog box informing them that a password expiration policy had taken effect and that they must change their passwords to proceed. At this point, participants in the experimental conditions saw password meters. Due to privacy concerns, we did not save participants' passwords, though we did save hashes of their original and changed passwords. We also collected the Levenshtein edit distances between these two passwords, the zero-order entropies, lengths, and the number of symbols from each character class. Participants actually changed their real passwords.

After changing their passwords, participants performed three subterfuge tasks that involved browsing the portal for information. After each of these tasks, they answered questionnaires in order to further convince them that this was the true purpose of the study and that the password change was not a planned part of the experiment. Upon completing all of the tasks, we compensated participants for their time.

We invited participants back to the laboratory two weeks later so that we could measure password memorability. We informed participants that they would be completing a followup survey on the portal, which required that they login again. We captured the same data with our proxy server as we did in the initial session. In addition to observing whether they were able to login, we also observed whether or not they had changed their passwords during the interim.[1] After this task, participants completed an exit survey that gathered qualitative data about their experiences. Finally, we debriefed them.

Participants

We recruited 51 participants with flyers around campus, as well as messages to various departmental mailing lists. Our only participation requirement was that participants had a university SSO account. We compensated participants with $20 after completing the first session, and an additional $25 after completing the second session.

During the experiment, one participant did not feel comfortable changing his password on a shared computer, another could not remember his initial password, and two others' data was lost due to proxy server difficulties. Thus, we were left with 47 participants in the initial session. Of these, fifteen were male (31.9%). Participants' ages ranged from 18 to the $56-65$ range, with a plurality of participants being in the $19-24$ range.[2] We observed no significant effects based on demographic factors and therefore do not discuss them further.

Results

Overall, we observed that both password meters yielded statistically significant differences when compared to the *control* condition. In this section we examine how the meters impacted password strength, what those impacts were, and whether usability was affected.

Password Strength

Prior to changing their passwords, participants' password strength did not significantly differ between the conditions. Across all conditions, passwords were an average of 46.7 bits strong ($\sigma = 10.15$). We performed Wilcoxon Signed Ranks tests to compare the bit strength of the previous and new passwords in each condition. After changing their passwords, the bit strength of participants in the *control* condition did not change significantly; the new passwords were 49.3 bits strong ($\sigma = 7.03$; $Z = 0.000$, $p < 1.0$).[3] However, we did observe statistically significant differences in both the *EM* and *PPM* conditions. In the *EM* condition, the zero-order entropy of new passwords increased to 60.8 bits ($\sigma = 16.00$; $Z = -3.180$, $p < 0.001$). In the *PPM* condition, the zero-order entropy of new passwords increased to 64.9 bits ($\sigma = 21.35$; $Z = -2.664$, $p < 0.008$). Since this effect was not present among users in the *control* condition, it is clear that the meters were responsible for nudging users towards stronger passwords. Figure 3 shows the feedback shown to participants in the experimental conditions.

We compared the changes in zero-order entropy between the three conditions using Mann-Whitney U tests. We observed that when compared to the *control* condition, passwords created in both the *EM* condition ($U = 63.000$, $p < 0.023$) and the *PPM* condition ($U = 54.500$, $p < 0.009$) contained significantly more entropy. However, there were no observable differences between our two experimental conditions ($U = 117.000$, $p < 0.678$). Thus, we reject H_0, accept H_1, and cannot accept H_2.

[1] We compared the cryptographic hashes, making it unnecessary for us to save or view participants' actual passwords.

[2] We did not collect exact ages, only ranges.

[3] All statistical tests were two-tailed unless otherwise noted.

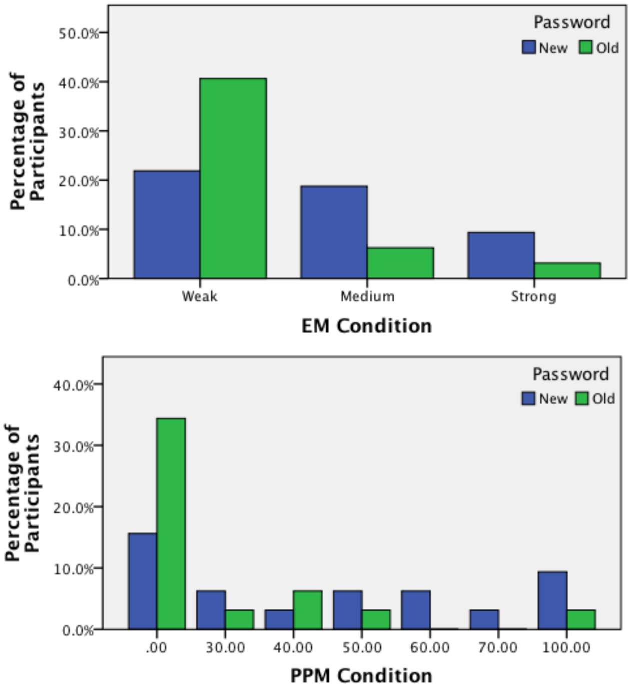


Figure 3. The feedback the meters would have showed for participants' old passwords and the feedback the meters actually showed for their new passwords.

Password Changes

We collected the following characteristics of participants' old and new passwords to see how passwords changed based on meter feedback (Table 2):

- Length
- Levenshtein edit distance
- Number of lowercase letters
- Number of uppercase letters
- Number of digits
- Number of symbols

We did not examine the new passwords of participants in the *control* condition because their strength did not significantly change. Likewise, the ways in which strength increased between the *EM* and *PPM* conditions did not observably differ. Thus, we merged the two experimental conditions and performed a Wilcoxon Signed Ranks test to compare the characteristics listed in Table 2, between participants' previous and changed passwords.

We applied the Holm-Sidak correction and found that with meters, passwords changed in three statistically significant ways. First, length increased from a median of 9.0 to 10.0 characters ($Z = -3.315$, $p < 0.0005$; one-tailed). Second, use of "special" symbols increased from zero to seven participants ($Z = -2.530$, $p < 0.006$; one-tailed). Third, lowercase letters increased from a median of 6.0 to 7.0 letters ($Z = -2.287$, $p < 0.011$; one-tailed). Thus, the meters motivated participants to create longer passwords through the inclusion of symbols and additional lowercase letters.

	Control (n=15)	**EM** (n=16)	**PPM** (n=16)
Length			
before	9.0 (6)	9.0 (5)	8.0 (6)
after	9.0 (4)	10.0 (9)	10.5 (11)
Lowers			
before	7.0 (10)	5.5 (7)	6.0 (9)
after	7.0 (10)	7.0 (11)	7.0 (15)
Uppers			
before	0.0 (0)	0.0 (5)	0.0 (3)
after	0.0 (5)	0.0 (3)	0.5 (2)
Digits			
before	2.0 (7)	3.0 (7)	2.0 (6)
after	2.0 (5)	2.0 (6)	2.0 (7)
Symbols			
before	0.0 (0)	0.0 (0)	0.0 (0)
after	0.0 (0)	0.0 (2)	0.0 (1)
Entropy			
before	46.53 (31.02)	46.53 (36.05)	47.08 (42.00)
after	51.70 (20.68)	60.10 (59.86)	59.45 (71.77)
Edit Distance	8.0 (11)	9.0 (12)	8.0 (13)

Table 2. Median password composition (and range) before and after the forced change in the laboratory study, as well as the Levenshtein edit distance.

Usability Concerns

Our results indicate that password meters—both traditional and those based on social pressure—can nudge users towards creating stronger passwords. However, nudging users to create stronger passwords may have drawbacks if users cannot remember them or choose to revert to weaker passwords. We measured whether our participants were still able to log in to their accounts two weeks after the experiment, as well as whether they had changed their passwords during the interim period.

Of our 47 participants, 40 re-authenticated and completed our exit survey. Of these, 10 (25.0% of 40) participants had since changed their passwords.[4] A chi-square test indicated that participants who were exposed to meters were no more likely to change their passwords than those in the *control* condition. Nine of the 10 participants who had subsequently changed their passwords reverted to their previous passwords. Four of these participants indicated that they did not want to remember an additional password. That is, their previous SSO password was used for other accounts, and as a result of our study, they needed to remember an additional password. Another 4 participants indicated that they had forgotten their new password, whereas the ninth participant said he was uncomfortable changing his password on a shared computer and therefore reverted to his previous password. Finally, the tenth participant who had changed his password indicated that he had done so because he had thought of an "even more secure" password.

The results of our exit survey indicate that while at least 19% of our participants reverted to their previous passwords, there is no evidence that this was because the meters nudged them into choosing overly burdensome passwords. Likewise, there is no evidence to suggest that

[4] Our results remain significant without these participants.

our results are confounded by the use of password managers, since no one reported using them. Participants in the *control* condition were just as likely to forget their new passwords or express frustration at the thought of having to remember yet another password. We believe this finding is a greater indictment of the burden of password expiration policies than of meters.

Discussion

While we found that password meters were effective, our laboratory experiment raised additional questions.

Changing vs. Creating

Participants changed passwords for existing accounts. It is unclear whether password meters have the same effect when users register new accounts. Given the rates of password reuse that have been documented in the literature [11] and our exit survey (22 out of 40 participants—55%—indicated they used their SSO passwords to protect other accounts), one might expect that many users will attempt to create an account with a reused password, rather than create a new password. The extent to which meters may mitigate this behavior is unclear.

Account Importance

Participants in our experiment used their actual passwords. As such, participants had clear incentives to choose very strong passwords—participants' original passwords were significantly stronger than the entropy of the bare minimum requirement, 43.6 bits ($p < 0.0005$; One-Sample Wilcoxon Signed Rank test). It is unclear whether users would expend similar effort in creating passwords for accounts they consider less important.

Meter Orientation

Participants in the *EM* condition saw a horizontal meter that took up minimal space. Participants in the *PPM* condition, however, saw a much larger vertical meter. This meter may have been more prominent, increasing the likelihood that participants noticed it (Figure 2). Thus, it is possible that the statistically insignificant difference between these two conditions was merely a lower bound, and that presenting a vertically oriented *EM* condition may produce a much larger effect size.

Sample Size

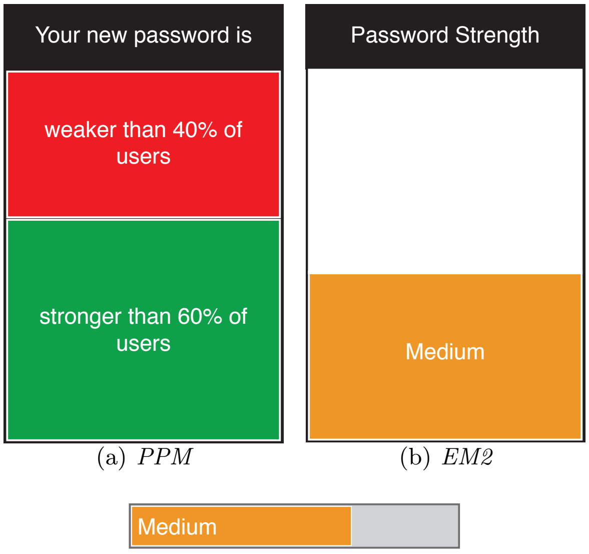
While the average entropy of passwords created under the *PPM* condition was greater than those created under the *EM* condition (64.91 bits vs. 60.76 bits), this difference was not statistically significant. It is possible that a much larger sample may have yielded statistically significant results. Thus, we cannot say whether differences in effects may exist between these two meters.

FIELD EXPERIMENT

Based on the open questions from our first experiment, we tested the following null hypotheses in the field:

H_{0a}: Passwords are not stronger when users see meters, when creating unimportant accounts.

H_{0b}: Changes to the orientation and text of password meters will not result in different passwords.



(a) *PPM* (b) *EM2*

(c) *EM*

Figure 4. The three experimental conditions.

Methodology

In this section, we describe our experimental conditions, protocol, and participants.

Conditions

We initially designed eight different experimental conditions to control for three different factors: meter orientation (horizontal or vertical), meter meaning (weak/strong or based on social pressure), and the choice between text and graphics to communicate that meaning. Thus, our intended conditions were as follows:

1. *Control*: No meter was displayed.
2. *EM*: A horizontal "weak" to "strong" meter, identical to the one in our laboratory experiment.
3. *EM2*: A vertical meter going from "weak" to "strong," but similar in area to the meter in the *PPM* condition.
4. *PPM*: A vertical meter depicting relative strength, identical to the one in our laboratory experiment.
5. *EM2NoTxt*: A vertical meter identical to the *EM2* condition, but with all text removed (i.e., the meter changed from red to orange to green).
6. *EMNoBar*: Words without graphics were displayed: "Your password is weak/medium/strong."
7. *EMNoTxt*: A horizontal meter identical to the *EM* condition, but with all text removed (i.e., the meter changed from red to orange to green).
8. *PPMNoBar*: Words without graphics were displayed: "Your password is stronger than *X*% of other users."

We ran a pilot on 200 participants, randomly assigned to the eight conditions. We observed no significant differences based on password entropy. Upon performing a power analysis ($\alpha = 0.05$, $\beta = 0.80$), we observed that we would need a sample size several orders of magnitude greater to yield significant differences between our latter four conditions. Thus, we concluded that if there were effects attributable to the text or the graphics, they were

Bit Strength (x)	PPM (stronger)	EM / EM2
x<=26.58	10%	Weak
26.58<x<=31.02	20%	Weak
31.02<x<=35.73	30%	Weak
35.73<x<=36.65	40%	Medium
36.65<x<=41.18	46%	Medium
41.18<x<=41.35	56%	Medium
41.35<x<=45.99	61%	Medium
45.99<x<=46.53	70%	Strong
46.53<x<=51.70	80%	Strong
51.70<x<=65.81	93%	Strong
65.81<x	100%	Strong

Table 3. Password strength intervals used to provide feedback in the field experiment.

so small as to be meaningless. As a result, we removed the latter four conditions, and recruited participants for the first three experimental conditions (Figure 4) and the *control* condition. Thus, we measured the effects of traditional password meters (the *EM* condition) and meters based on social navigation (the *PPM* condition), while controlling for meter orientation (the *EM2* condition).

We calibrated the meters with the entropy distribution found in the RockYou dataset (Table 3). We reasoned that the entropy distributions would be similar since they were collected without minimum requirements and neither account was likely considered "important."

Protocol

As in our first experiment, we did not want participants to know we were studying passwords. To accomplish this, we added an account creation page to a website being used for another, unrelated study. In that study, participants visited the website of a fictitious startup that was beta testing an Android application in order to gather behavioral data on smartphone application pricing [9]. This website was privately registered and could not be linked with us or our institutions. Participants in that other study had no reason to disbelieve our explanation. For this study, we added a page to that website so participants could create accounts to register for a private beta. This page featured password and password confirmation fields. We did not list or enforce any minimum password requirements. We randomly assigned participants to one of our four between-subjects conditions. We collected usernames and passwords, as well as instrumented the page to record the amount of time it took each participant to type a password.

We intentionally did not tell participants when or if we would be contacting them again, because we did not want to bias them towards writing their passwords down or otherwise expending additional effort on remembering them; we wanted the registration and subsequent authentication tasks to be as realistic as possible in order to maximize ecological validity. Two weeks after registering on our website, we sent each participant a message containing a link to a login page. We explained that upon successfully logging in, they would receive a $0.50 bonus payment for their time, as well as see whether they qualified for the beta test. On the login page, we allowed participants unlimited login attempts, but did not allow them to reset or recover their passwords. The reason for this was that we wanted to observe the number of attempts participants would make when forced to recall their passwords. Upon logging in, we informed them that they did not qualify for the beta test.

Finally, after a month had passed, we emailed participants to inform them that they had taken part in a study on passwords.[5] We included a link to an exit survey and offered a $2 payment for successful completion. We asked participants how they created the passwords used in this experiment, whether they used these passwords for other accounts, and how strong they believed these passwords were compared to their other passwords. Thus, before accessing the survey, we asked them to login again to ensure that the password that was the subject of the survey was fresh in their minds. Because the purpose of the login task was to prime participants, rather than re-examine password memorability, we allowed participants to receive forgotten passwords via email.

Participants

We recruited participants using Amazon's Mechanical Turk. Our only requirements were that participants be over 18 years of age and in the U.S. Because this experiment was run in conjunction with another experiment that was focused on Android users, all of our participants were also Android users. A total of 541 participants created passwords in the first part of our experiment. While we cannot identify the precise demographics of the subset of subjects who participated in this study, 61.3% of the 763 participants in the Android study were male, with an average age of 29 ($\sigma = 9$) years [9].

Results

When participants created passwords for unimportant accounts, we observed no effects that could be attributable to the presence of the meters. This contrasted with our first experiment, in which participants who were shown meters chose significantly stronger passwords when changing the passwords for important accounts. In this section, we present our results in terms of password strength, memorability, and our exit survey results.

Password Strength

We found no statistically significant differences between any of our conditions with regard to bit strength, length, or composition: we cannot reject either H_{0a} or H_{0b}. Overall, participants' passwords had a median bit strength of 41.4 and were a median of 8 characters long.

We were concerned that data from participants who failed to subsequently login may skew our data, since we cannot know whether they forgot their passwords or did not take the task seriously. For example, some participants may have entered gibberish if they never expected to login again. However, we found no evidence of this; the length and entropy of participants' passwords did not

[5] We excluded participants who never attempted to login.

	Control	**EM**	**EM2**	**PPM**
All participants ($n = 541$)				
n	120	141	144	136
Length	8.0 (14.0)	8.0 (14.0)	8.0 (14.0)	8.0 (15.0)
Entropy	41 (78)	41 (106)	44 (102)	44 (98)
Time	4.5 (34.0)	6.0 (159.0)	6.0 (87.0)	7.0 (38.0)
Returning participants ($n = 331$)				
n	76	84	81	90
Success	55 (72%)	57 (68%)	61 (75%)	71 (79%)
Successful participants ($n = 244$)				
n	55	57	61	71
Tries	1.0 (13.0)	1.0 (11.0)	1.0 (8.0)	1.0 (6.0)

Table 4. Length, entropy, and creation time medians and ranges. Next, whether participants could log in two weeks later and the median number of attempts it took them.

significantly change based on whether or not participants attempted to log in, or even whether they were successful. Nonetheless, Table 4 depicts the median lengths, bit strengths, and sample sizes across the four conditions.

We observed no statistically significant differences with regard to strength metrics between the three experimental conditions and the *control* condition. We hypothesize that this may be partially due to unexpectedly strong passwords across all of our conditions (including the *control*). Two-thirds of our participants employed multiple character classes and relatively long lengths, despite the lack of minimum strength requirements. In fact, only 33.3% of our 541 participants used a single character class (5.0% used only numbers, while 28.3% used only lowercase letters) and only 24 (4.4% of 541) participants created passwords that were shorter than six characters.

We performed Levene's Test for Equality of Variances to compare the entropy distributions between the *control* and the three experimental conditions. We observed no significant differences (*EM*: $F = 0.173$, $p < 0.678$; *EM2*: $F = 0.008$, $p < 0.927$; *PPM*: $F = 0.137$, $p < 0.712$), which led us to question whether our null result was due to participants simply not noticing the meters. Since the meters were updated in realtime as participants typed, we hypothesized that if the meters were noticed, participants may interrupt their typing, which could result in significantly longer password creation times. Indeed, Mann-Whitney U tests comparing the creation times between the experimental conditions and the *control* were statistically significant (*EM*: $U = 6786.00$, $p < 0.006$; *EM2*: $U = 6820.00$, $p < 0.003$; *PPM*: $U = 6038.00$, $p < 0.0005$). Thus, our empirical data suggests that while participants noticed the meters, their resulting passwords were ultimately unaffected by them.

Password Memorability

Two weeks after participants created passwords, we asked them to return to our website. This required logging in, though no password recovery or reset mechanism was available. We examined whether any of the conditions significantly differed with regard to password memorability. Table 4 depicts the number of participants who attempted to login, those successful, and the median number of attempts it took them.

We observed no significant differences with regard to the proportion of participants in each condition who either attempted or succeeded at logging in. Overall, 331 (61.2% of 541) participants attempted the task and 244 (73.7% of 331) succeeded, the majority of which did so on their first attempt (60.2% of 244). The 87 participants who gave up made a median of 3 attempts. We observed no significant differences with regard to password strength between participants who were successful and those who were not; password strength was not correlated with memorability ($r = -0.030$, $p < 0.591$).

We were concerned that the use of browser-based password saving features may bias our results. To check for this, we measured the amount of time participants spent typing their passwords. We found evidence that only 3.2% (95%CI: [1.3%,6.5%]) of participants used these features, and therefore they did not influence our results. Thus, while the meters did not nudge participants into choosing significantly stronger passwords over those in the *control* condition, participants who viewed meters were no more likely to forget their passwords either.

Survey Results

A primary goal in both of our experiments was to maximize ecological validity during the password creation and login phases by making these tasks required steps to complete larger subterfuge tasks. Thus, up until this point, we did not reveal the true purpose of the experiment. Based on the divergent results between our laboratory and field experiments, we ended the deception by inviting our field experiment participants to answer an exit survey regarding their password choices. We ensured that participants knew the password about which we were asking by forcing them to login again, but allowed them to recover forgotten passwords by email. Of our 331 participants who attempted to log in, 218 completed this survey.

Similar to our first experiment, we observed widespread password reuse among participants: 132 (63.8% of 207) reported using their passwords elsewhere.[6] Reuse rates did not significantly differ between the four conditions, indicating that the meters did not observably nudge participants towards creating new passwords. We hypothesize that 63.8% (95%CI: [56.8%, 70.3%]) represents a lower bound, as some participants may not have admitted that they knowingly engaged in poor security practices. Nonetheless, we observed several significant correlations that corroborate password reuse. For instance, participants who reused passwords were likely to spend less time typing passwords during the first phase of this experiment ($r = -0.219$, $p < 0.005$). Participants who claimed to reuse passwords were more likely to remember them during the second phase ($r = 0.245$, $p < 0.0005$), and less likely to use the password recovery feature to access the exit survey ($r = -0.323$, $p < 0.0005$).

[6] Not all 218 participants answered every question.

Not only did participants reuse existing passwords, but they knowingly reused weak ones. We asked participants to rate the strength of their password relative to their other passwords using a 5-point Likert scale (from "much stronger" to "much weaker," with "similar" as the neutral option). Only 37 participants (17.9% of 207) responded that their study passwords were either "stronger" or "much stronger" than their other passwords. Likewise, a Wilcoxon Signed Ranks test indicated that participants' observed experimental passwords were significantly shorter than their self-reported longest passwords ($\mu_{observed} = 8.57$, $\mu_{reported} = 12.34$, $Z = -9.217$, $p < 0.005$). However, we found that reused passwords were not observably weaker than the passwords of those who claimed not to have reused passwords. Thus, the extent to which password reuse impacts strength remains unclear. We believe that effects stemming from participants' perceptions about the unimportance of the website outweighed any effects relating to the meters or their choice to reuse existing passwords; when passwords were reused, weaker existing passwords were employed.

Only 21 (12.7% of 156) participants remembered seeing the meters. Others acknowledged that if meters were shown, they would have labeled their passwords as weak:

- *"I'm sure it would have said it was weak."*
- *"When I use this generic web password on other sites, their meters always say it is weak."*
- *"This one is usually yellow...I have other options if I feel I need a very strong password, like for banking."*

Thus, the results of our field experiment suggest that when password meters are shown when creating new accounts on websites that users consider unimportant, the meters are unlikely to influence password strength.

LIMITATIONS AND CONCLUSION

Our main contribution is in showing how password creation behaviors are heavily dependent on context. Some may be quick to charge this as obvious; while our results may not be very counter-intuitive, we point out that they suggest that current practice at many major websites then defies the obvious. For example, one of our findings is that password meters do not yield much improvement in helping users choose passwords for unimportant accounts, yet they are very commonly deployed in such contexts. Equally, where meters make a difference—password changes for important accounts—they are less often seen. Thus, practice at real sites appears to be very far from what our results dictate. This indicates a real opportunity for improvement.

We tested the impact of two variables on password meter effectiveness: creating a new account vs. changing the password on an existing account, and doing so on important vs. unimportant accounts. Because we only performed two experiments, rather than the four needed to exhaust the space, we do not know the extent to which each variable independently influenced behavior. Likewise, because each experiment was performed at different times, with different protocols, using different sample populations, we cannot perform statistical comparisons between the two and instead only report on each's respective results.

As with any study, ecological validity is hard to ensure. While we made a concerted effort in both experiments to mask our primary interest in participants' passwords, we cannot be absolutely sure that no participants saw past the deception. That said, we observed no evidence that our results were due to the Hawthorne effect; if subjects created stronger passwords solely because they believed that was what we wanted, we would not have observed significant differences between conditions in the laboratory (i.e., all passwords would have been stronger, regardless of the randomly-assigned condition).

In both experiments, majorities of participants reported reusing passwords: 55% in the laboratory and 63.8% in the field. Only when laboratory participants were forced to change their passwords while viewing meters did they choose stronger passwords (though 22.5% reverted to their old passwords, regardless of the meters, because they did not want to remember another password). It is unclear whether the meters impacted password reuse behaviors. For instance, it is possible that when creating a new password for a high-risk account, users may still reuse a password, but may be nudged into reusing one of their stronger existing passwords.

One interpretation of our results is that presenting a password meter at the time of registration is too late, because users already know which of their existing passwords they plan to reuse. However, significant improvement is achieved when users are creating new passwords. This suggests that password meters not associated with account registration pages (e.g., an informational website) might have considerable influence. The minority of users who seek out such feedback are probably far more amenable to influence than the average user (though at the same time, such users are probably more likely to already understand what constitutes a strong password).

Some motivation for password meters seems guided by the belief that users do not understand when their passwords are weak. The results of our study draw this belief into question. We found that, in many cases, participants knowingly chose weak passwords. At least in the case of unimportant accounts, they demonstrated an understanding that the password they used was not merely being reused, but also weak. Weakness was not a problem of which they were unaware, but one of which they were aware but insufficiently motivated to fix.

ACKNOWLEDGMENTS

We would like to thank Adrienne Porter Felt, Maritza Johnson, John Manferdelli, Daniele Perito, and David Wagner for providing feedback on drafts of this paper, and members of the Laboratory for Education and Research in Secure Systems Engineering (LERSSE) for providing feedback on the various stages of the laboratory experiment. This work was supported by the Intel Science and Technology Center for Secure Computing, as well as the Canadian NSERC Internetworked Systems Security Network (ISSNet).

REFERENCES

1. John the Ripper. http://www.openwall.com/john/. Accessed: January 4, 2013.
2. Adams, A., and Sasse, M. A. Users are not the enemy. *Communications of The ACM 42* (December 1999), 40–46.
3. Alexa Internet, Inc. Alexa top 500 global sites. http://www.alexa.com/topsites/global;1, 2012. Accessed: April 23, 2012.
4. Besmer, A., Watson, J., and Lipford, H. R. The impact of social navigation on privacy policy configuration. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, ACM (New York, NY, USA, 2010).
5. Bonneau, J. The science of guessing: analyzing an anonymized corpus of 70 million passwords. In *IEEE Symp. Security and Privacy* (2012).
6. Castelluccia, C., Duermuth, M., and Perito, D. Adaptive password-strength meters from markov models. In *Proceedings of the Network & Distributed System Security Symposium (NDSS), San Diego, CA* (2012).
7. D.A. Norman. The Way I See It: When security gets in the way. *Interactions 16*, 6 (2009), 60–63.
8. Dell'Amico, M., Michiardi, P., and Roudier, Y. Password strength: An empirical analysis. In *IEEE INFOCOM 2010*, IEEE (March 2010), 1–9.
9. Egelman, S., Felt, A. P., and Wagner, D. Choice architecture and smartphone privacy: There's a price for that. In *The 2012 Workshop on the Economics of Information Security (WEIS)* (2012).
10. Egelman, S., Molnar, D., Christin, N., Acquisti, A., Herley, C., and Krishnamurthi, S. Please continue to hold: An empirical study on user tolerance of security delays. In *Proceedings (online) of the 9th Workshop on Economics of Information Security* (Cambridge, MA, June 2010).
11. Florêncio, D., and Herley, C. A large-scale study of web password habits. In *Proceedings of the 16th International Conference on the World Wide Web*, ACM Press (New York, NY, USA, 2007), 657–666.
12. Forget, A., Chiasson, S., Van Oorschot, P., and Biddle, R. Improving text passwords through persuasion. In *Proceedings of the 4th symposium on Usable privacy and security*, ACM (2008), 1–12.
13. Herley, C., and van Oorschot, P. C. A research agenda acknowledging the persistence of passwords. *IEEE Security & Privacy 10*, 1 (January/February 2012), 28–36.
14. J. Yan and A. Blackwell and R. Anderson and A. Grant. Password Memorability and Security: Empirical Results. *IEEE Security & Privacy* (2004).
15. Komanduri, S., Shay, R., Kelley, P. G., Mazurek, M. L., Bauer, L., Christin, N., Cranor, L. F., and Egelman, S. Of Passwords and People: Measuring the Effect of Password-Composition Policies. In *CHI '11: Proceeding of the 29th SIGCHI Conference on Human Factors in Computing Systems*, ACM Press (New York, NY, USA, 2011).
16. Kuo, C., Romanosky, S., and Cranor, L. Human selection of mnemonic phrase-based passwords. In *Proceedings of the second symposium on Usable privacy and security*, ACM (2006), 67–78.
17. R. Morris and K. Thompson. Password Security: A Case History. *Comm. ACM* (1979).
18. Schechter, S., Herley, C., and Mitzenmacher, M. Popularity is everything: A new approach to protecting passwords from statistical-guessing attacks. In *Proc. HotSec'10* (2010).
19. Shay, R., Komanduri, S., Kelley, P. G., Leon, P. G., Mazurek, M. L., Bauer, L., Christin, N., and Cranor, L. F. Encountering stronger password requirements: user attitudes and behaviors. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, ACM (New York, NY, USA, 2010), 2:1–2:20.
20. Sotirakopoulos, A. Influencing users password choice through peer pressure. Master's thesis, University of British Columbia, 2011.
21. Thaler, R., and Sunstein, C. *Nudge: Improving decisions about health, wealth, and happiness.* Yale University Press, New Haven and London, 2008.
22. Ur, B., Kelley, P. G., Komanduri, S., Lee, J., Maass, M., Mazurek, M. L., Passaro, T., Shay, R., Vidas, T., Bauer, L., Christin, N., and Cranor, L. F. How does your password measure up? the effect of strength meters on password creation. In *Proceedings of the 21st USENIX Security Symposium* (2012).
23. W. E. Burr, D. F. Dodson W. T. Polk. Electronic Authentication Guideline. In *NIST Special Publication 800-63* (2006).
24. Weir, M., Aggarwal, S., Collins, M., and Stern, H. Testing metrics for password creation policies by attacking large sets of revealed passwords. In *Proc. ACM CCS '10* (2010).
25. Weir, M., Aggarwal, S., de Medeiros, B., and Glodek, B. Password cracking using probabilistic context-free grammars. In *The 30th IEEE Symposium on Security and Privacy*, IEEE (2009), 391–405.