Hybrid Password Meters for More Secure Passwords - A Comprehensive Study of Password Meters including Nudges and Password Information

Verena Zimmermann
Technische Universität Darmstadt, Germany
zimmermann@psychologie.tu-darmstadt.de

Karola Marky
University of Glasgow, Glasgow, Scotland
Technische Universität Darmstadt, Germany
Keio University, Yokohama, Japan
karola.marky@glasgow.ac.uk

Karen Renaud
University of Strathclyde, Glasgow, Scotland
Rhodes University, RSA
University of South Africa, RSA
Abertay University, Scotland
karen.renaud@strath.ac.uk

ABSTRACT

Supporting users with secure password creation is a well-explored yet unresolved research topic. A promising intervention is the password meter i.e. providing feedback on the user's password strength *as and when* it is created. However, findings related to the password meter's effectiveness are varied. An extensive literature review led us to the conclusion that, besides providing password feedback, effective password meters often also include: (a) feedback nudges to encourage stronger passwords choices, and (b) additional password guidance. A between-subjects study was carried out with 645 participants to test nine variations of password meters with different types of feedback nudges exploiting various heuristics and norms. This study explored differences in resulting passwords: (1) actual strength, (2) memorability, and (3) user perceptions. The study revealed that password feedback, in combination with a feedback nudge and additional guidance, labelled a *hybrid password meter*, was generally more efficacious than either intervention on its own, on all three metrics. Yet, the *type* of feedback nudge targeting either the person, the password creation task, or the social context, did not seem to matter much, the meters were nearly equally efficacious. Future work should focus on the short- and long-term effects of hybrid password meters in real-life settings to confirm the external validity of these findings.

Author Keywords

Authentication; Password Meter; Nudge; Password Creation; User-Centered Design

INTRODUCTION

Passwords are still the most commonly used authentication mechanism [101]. Even in areas where biometrics are on the rise, such as fingerprint authentication on the mobile phone, passwords are still a widely used fallback mechanism [108]. Furthermore, several studies that investigated authentication mechanisms concluded that passwords, as an authentication mechanism, are likely still to be around for the foreseeable future [49, 13, 86, 51, 97].

The shortcomings of passwords are well-researched: they are susceptible to keystroke logging, vulnerable to phishing attacks, and also to guessing attacks [49]. A significant problem arises from the potential mismatch between the user's password strength perceptions and the actual strength of the password [110, 90]. Another significant factor contributing towards their flaws is the behaviour of the password owner. The increasing cognitive load posed by managing multiple strong, unique passwords [29, 102] leads to the use of coping strategies, such as the use of weak and memorable passwords [40, 113, 110], and the reuse of passwords across multiple accounts [94, 118, 74]. Coping strategies and misconceptions compromise password security, and make it easy for attackers to breach several accounts using one leaked password [17]. Because password security depends not only on technical measures, but also on users' password creation and management strategies, it is important to explore ways to encourage the creation of secure, yet memorable, passwords.

Existing strategies such as system-generated passwords, regular password expiry and enforced password policies enhance password security from a technical perspective. However, those strategies tend to neglect human factors that impact password security in the longer term [39]. In particular, system-generated passwords are generally not memorable, leading users to write them down [72, 3, 128]. Moreover, the security benefit of regular password expiry is questionable [17, 127], is a source of frustration to users [87], and leads to the use of slight variations of existing passwords [43]. This maximises predictability and weakens the mechanism [17]. Password policies have been shown to be less effective in increasing

Submitted to Behaviour & Information Technology

password strength than initially anticipated [120], and also negatively impact memorability and usability [50].

Password meters are a promising user-centered strategy to improve password security [109, 100]. Password meters, as depicted in Figure 1, provide users with feedback, in real time, reflecting the strength of the password they are creating, without enforcing minimum strength. The strength-related feedback can be displayed in the form of a feedback bar that fills as strength increases, as shown in Figure 1. Textual feedback, visual elements or password scores can also be provided. Sometimes, password meters also provide suggestions for improving the strength of the password during password creation. Password meters are frequently deployed e.g. on eBay, Google, Facebook, and Twitter [22, 31, 111]. Van Acker *et al.* [115] found that password meters were used by about a third of the top 250 Alexa domains. Password meters can address the strength issues associated with passwords in three ways. *First*, by providing strength feedback, password meters can bridge the gap between the users' perceptions of their password's strength and technical password strength. *Second*, the feedback may motivate users to increase password strength. *Third*, by encouraging stronger passwords, users might be persuaded to deviate from their usual password creation routines and this could potentially prevent password reuse. Indeed, a number of studies have found that password meters do increase the strength of user-created passwords [109, 84, 116]. While password meters might not be able to solve the password memorability problem, related work suggests that the passwords created when password meters were displayed are *as memorable as* passwords created in their absence. [27, 85, 111].

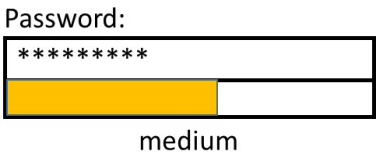

Figure 1. Exemplary depiction of a password meter with a feedback bar including a simple nudge (color-coding).

Instead of *forcing* users to change their passwords, meters follow a softer approach [73, 90]. Password meters often include nudges [107]: small interface tweaks that aim to encourage users to increase their password strength without forcing them or limiting their options in terms of choosing the password they want. Instead, the nudges target the password creator's automatic cognitive processes [107] by activating biases, heuristics and norms. Examples include the use of color-coding [109] evoking the learned connection between the color red and bad/insecure/unfavorable or the color green and good/secure/favorable (See Figure 1), the use of social norms eliciting perceived peer pressure [84, 27], or fear appeals [116].

Even so, password meters have not enjoyed unanimous success. Researchers have identified a number of inconsistencies in modelling password strength with different strength estimations or cracking algorithms [114, 37]. This kind of variability in the modelling of strength or guessability of passwords could easily lead to confusion among users having similar passwords rated differently across accounts. Furthermore, while password meters were successful in encouraging the creation of stronger passwords in some studies [109, 84, 116], this effect was not replicated in other studies or in other password meter conditions within the same studies [27, 38, 116, 80]. For example, Vance *et al.* [116] found an interactive password meter to be successful while a static one was not. In a longitudinal field study, a password meter did not significantly increase password strength [80], whereas linking password strength to password expiry was successful [84].

The question remains: "*What makes password meters effective?*" To answer this higher-level question, a systematic literature review was conducted as a first research step. Analysing 42 publications that described 108 different password meter variants revealed that password meters seem to be effective when they include: (a) password strength feedback, (b) some kind of visual feedback nudge (e.g. a social comparison), and (c) additional password creation guidance. This combination is termed a *hybrid password meter* in this paper.

The review led to the hypotheses H_1 - H_3 that hybrid password meters are more effective in encouraging stronger, longer and high entropy passwords than non-hybrid password meters. Via the following research questions (RQ), we also aimed to explore whether: (RQ_1) the type of visual feedback nudge influenced password creation, (RQ_2) their impact on memorability, and (RQ_3) user perceptions of password meters. The research questions and hypotheses guiding the research are graphically depicted in Figure 2. To explore the hypotheses, a between-subjects online study with $N = 645$ users was conducted in a second research step. It compared six password meter variations, based on those proposed by Ur *et al.* [109], to the original password meter and two control conditions. As control conditions, variants of Ur *et al.'s* [109] study that only contained password guidance or a feedback bar were used. The hybrid nudge variations included a motivation and fear appeal nudge targeting the person, a reciprocity and password lifetime nudge to target the password creation context, and a descriptive and normative norm nudge targeting the social context. The study revealed that, overall, the hybrid nudge variations were more successful in encouraging password strength, as compared to the two control conditions. This finding was confirmed by the positive user perceptions and the high memorability rates of passwords created in the presence of a hybrid password meter. Yet, it seems that the specific heuristic or norm the feedback nudge targeted did not make a significant difference. In particular, it does not seem to matter whether the nudge targets the person, the person within the password creation context, or the person's social context. Only in one case did the study reveal a significant difference between hybrid nudge variations. Future work should validate the findings in a field setting and also assess their long-term influence.

Contribution

To summarize, the contribution of this research is threefold:

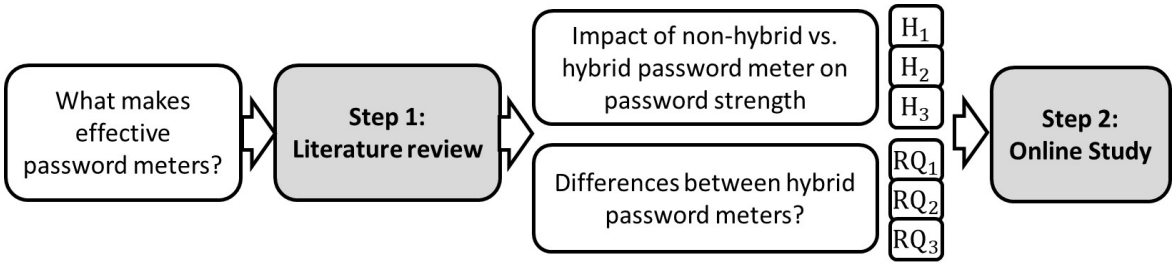

Figure 2. Process of identifying and exploring hypotheses and research questions.

- First, an extensive literature review on password meters shines light on the factors that contribute to their effectiveness.
- Second, an online study with 645 participants that systematically compares the influence of the type of feedback nudge used by hybrid password meters on the passwords participants created.
- Third, the comparison considers nudges specifically designed to target the person, the password creation context, or the social context, revealing nudge efficacy within the specific context of use.

The remainder of the paper is structured as follows. The next section provides some background information on the criteria for nudging. Afterwards, the process and outcome of the literature review is detailed. The assumptions and questions derived from the literature review outcome are detailed in section *research questions and hypotheses*. After describing the method of the study to analyse the questions, the *results* are presented. Finally, the discussion of the results is followed by the conclusion.

BACKGROUND: CRITERIA FOR NUDGES

Password meters often rely on "nudges". The term was coined by Thaler and Sunstein, in 2008, in their seminal book [107]. Nudges are small tweaks in the "choice architecture" used to encourage users to veer towards more secure passwords in this study. They have proven potential to positively impact choice in a variety of areas such as encouraging healthy behaviors [58], or energy-saving actions [5, 77]. In the digital space, including cybersecurity, initial research highlights the promise of nudging [1, 16], not only in terms of password creation [84, 24], but also in terms of encouraging privacy-preserving choices [18].

To count as a nudge, an intervention has to satisfy four criteria:

1) Predictability. A nudge should alter a person's behavior in a predictable way [107]. Therefore, the intended outcome and the intervention to achieve that aim should be carefully and deliberately chosen. For example, a proven password nudge might be chosen to support users in generating secure yet memorable passwords by nudging them towards the use of passphrases.

2) Preservation of Choices. Nudges should not limit the number of pre-nudge options [107]. For example, password policies or blacklists do not satisfy this criterion because they restrict options.

3) Equality of costs. Nudges do not considerably influence the options' "weight", e.g. by providing financial incentives for the secure option [107, 47]. For example, paying users or providing extended service functionality for a secure password would render the options unequal in terms of cost.

4) Automatic cognitive processes. Nudges should target automatic cognitive processes, such as biases or heuristics [15, 47, 44]. An example in the password creation context is the use of a social comparison nudge to invoke social norms.

In addition to the four criteria, the application of nudges should be carefully considered from an ethical perspective leading to a fifth requirement:

5) Ethical deployment. Because users might not be aware of the influence of interventions that target automatic information processing, it is important to make the nudge transparent to the nudgee [81, 45]. Sunstein *et al.* [106] emphasize that nudgees should agree that the nudge is warranted. To assure this, they suggest that nudges be open and transparent. This might be achieved by making use of informative statements and visualizations, to ensure that the nudge is deployed ethically.

LITERATURE REVIEW

A literature review was conducted to shine light on the reasons for efficacy differences reported by password meters studies. The aim was to: (1) construct an overview of a wide range of studied password meters, and (2) to reveal design differences in the meters and in the empirical evaluations thereof that might explain differences in study outcomes. The exploratory research question could be phrased as "*What makes password meters effective?*" A password meter can be defined as "*any tool that provides users with feedback on their password's strength while they are creating it*". The form of the feedback is not restricted and could be a bar, a visual or textual element. We reviewed feedback-only password meters as well as those that contained additional elements such as password creation information. Appendix A describes all identified password meters, along with their design elements.

The literature review on the effectiveness of different password meters was conducted within a time frame from 2000 to

2020[1]. The year 2000 was chosen as starting point as it marks the beginning of research in the area of usable security and privacy as, for example, encouraged by Adams and Sasse [2], and Whitten and Tygar [123]. The literature search process is depicted in Figure 3.

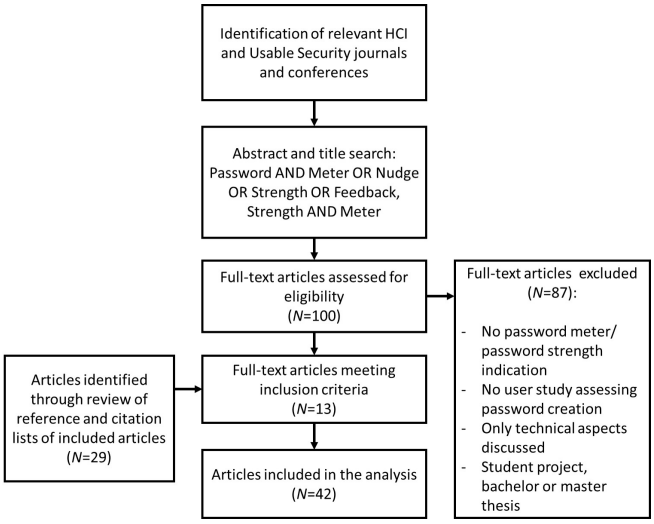

Figure 3. Prisma of the literature search process.

Search Criteria

Similar to the literature review conducted by Caraban *et al.* [16], this search included the top ten conferences and journals in the area of human-computer interaction, as rated by Google Scholar. Additionally, the following top conferences and journals in the field of usable security and privacy potentially including user-centered and password-related research were chosen: the Symposium on Security and Privacy (S&P), the Annual Computer Security Applications Conference (ACSAC), USENIX Security, the Conference on Computer and Communications Security (CCS), the Symposium on Usable Privacy and Security (SOUPS), the Network and Distributed System Security Symposium (NDSS), the (European) Workshop on Usable Security (Euro)USEC, the conference Passwords, and the journals Human-Computer Interaction (HCI), Computers in Human Behavior, and Information and Computer Security (ICS).

A complete list of the searched conferences and journals can be found in Appendix A .1. The search terms include a number of variations of terms in the title and abstract: “Password AND Meter”, “Strength AND Meter”, “Password AND Nudge”, “Password AND Strength” and “Password AND Feedback”.

Exclusion Criteria

For the purpose of this research, publications were only included if they described one or more password meters, and their impact on password strength during creation and memorability subsequently, and were evaluated in some kind of user study. In line with the definition provided in the previous section, we considered a password meter to be any tool that provided users with an indication of password strength. This including those that provided information or suggestions for stronger password creation. Publications describing concepts that had not yet been evaluated, or those that only described technical aspects such as the algorithms underlying a password meter, were excluded from the analysis [2].

Bachelor and Masters theses, as well as research that had not been peer-reviewed, were excluded. At the end of this process, 29 additional articles were included to support analysis.

After Exclusion

The application of the search criteria resulted in 96 publications. After applying the exclusion criteria, 13 publications remained for further analysis. Appendix A provides an overview of the number of publications identified in the different conference proceedings and journal databases.

Forward and Backward Searches

For each of the 13 relevant articles, an additional forward and backward search was carried out to identify potentially relevant research the authors referred to, or future studies that cited the publication. By so doing, we extended the search space beyond the already searched journals and conferences. For the forward and backward search, the same exclusion criteria were applied.

Analysis and Outcome

The 42 publications were examined in terms of the applied study design, the features of the applied *password meters*, and the *outcome* of the evaluation.

Study Context

Most of the studies (N = 27) used an artificial study context, asking users to role play creating an account. These studies were mainly conducted online, using different crowd-sourcing platforms. A total of 10 studies[3] either studied real account passwords in some kind of field study or made users believe they were creating a password for a real account. The majority of studies (N = 29) used a between-subjects design, that is, each participant interacted with one password meter. A total of five studies had users interact with, and compare, multiple meters in a within-subjects design. Two studies used a pre-post design with one password meter. Furthermore, of the 34 studies using either within- or between-subjects design, four also conducted a pre-post comparison, thus combining different measures. Participants were mainly recruited from university students and staff (N = 15), or from various crowd-sourcing platforms, such as Amazon Mechanical Turk, Lancers or Prolific (N = 13). Three studies included users of a certain app or website. Five studies did not explain how they recruited their participants.

The Password Meter

Overall, the 36 studies described and evaluated 108 different password meter variants using different types of feedback nudges, ranging from one variant to a maximum of 14 variants

[1] All articles available online in March 2020

[2] For a comparison and discussion of the technical accuracy of strength meter estimates see Golla and Dürmuth [37]

[3] One study consisted of two sub groups.

within one study. These could be classified as shown in Table 1. Most of the password meters ($N = 41$) made use of a colored feedback bar filling with increasing password strength combined with textual feedback, such as "weak" or "strong" (see Figure 1). Sometimes the bar took a slightly different shape, e.g., a dial [21].

Password meters in the second category used only textual information and suggestions, without a visual depiction. This included simple, but also extensive feedback [109], or character insertion suggestions for increasing password strength [89].

A total of 13 variants aimed to increase user motivation by making use of gamification elements. These ranged from password scores (e.g. [71]) to game-like environments. For example, Aljaffan *et al.* [4] created a radar chart that visualized certain password risks that users should eliminate by adapting their password. Furnell *et al.* [32] developed a game in which users had to include suggested characters to achieve a high password score under time pressure. Other approaches used motivational statements [25] or a visualization of a bunny dancing as password strength increased [111].

Also frequently used were social comparisons ($N = 12$) in which the user's own password was visually or textually compared to other users' passwords. Nine variants provided a strength indicator in the form of the time needed to crack the password (e.g. [53]) or the guessability of the next character [56]. These approaches were often labelled as fear appeals (e.g. [116]).

Another eight variants aimed to invoke a certain emotion by making use of affective messages [96] or emoticons [41].

Two variants linked password strength to password lifetime, both in real university accounts [8, 84].

Seitz *et al.* [91] tested whether the decoy effect could be exploited to encourage the selection of stronger passwords. Finally, Kim *et al.* [55] provided password feedback across multiple accounts addressing password reuse and account sensitivity.

Fear appeals, social comparisons and other approaches sometimes also made use of feedback bars but were sorted into their specific category as the primary purpose of the bar was to e.g. facilitate the social comparison or visualize scores within a gamification approach. Thus, password meters were categorized according to their primary purpose as described in the article.

In 45 variants, the password meter did not only provide a rating of current password strength, but also included further information - or links to information (e.g. [54]) - on what makes a strong password. This could be static textual information of password creation, e.g. [116, 33], or dynamic suggestions for improving the password strength such as "*Consider making your password longer*", e.g. [111, 109].

The password meters were compared to a control condition without a password meter in 17 studies. This included studies that used a pre-post comparison. A total of 24 studies compared the password meters to other experimental conditions involving different password meter variants. Eight studies also used other intervention types as a comparison, such as password policies or system-generated passwords (e.g. [100]). Finally, one study did not use a comparison [126], and one study analysed differences between people that were encouraged to change their password due to the password meter feedback, as compared to those who were not [21]. The number of studies exceeds 36 as multiple comparisons were possible within one study.

Type of Meter	Example	N
Feedback Bar and Textual Feedback	colored bar filling with increasing password strength	42
Textual Feedback and Suggestions	insertion suggestions	21
Gamification and Motivation	password scores, dancing bunny	13*
Social Comparison	comparison to average users' password	12
Fear Appeal	password guessing time in online attack	9*
Affect and Emotion	emoticons, affective messages	8
Password Lifetime	password lifetime increasing with password strength	2
Decoy Effect	suggestion of 2 alternatives	1
Feedback Across Accounts	multi-password feedback	1

Table 1. List of the type of password meters used in the analysed studies. *One study combined gamification elements with a fear appeal.

The Outcome

The studies' outcomes were measured on a variety of different measures including the estimated number of guesses based on different models ($N = 14$), different kinds of password scores ($N = 13$), password entropy ($N = 7$), password length ($N = 15$), or password composition ($N = 12$). Furthermore, eight studies measured the perceived influence on password creation, e.g. in terms of security (awareness) [95] or the perceived increase in strength [85]. Two studies did not collect any strength-related measures [4, 103]. The numbers of measures exceed the numbers of studies as many studies used multiple measures.

Of the 36 studies, $N = 19$ reported some positive and significant impact of at least one of the password meter conditions tested and $N = 8$ reported a positive increase in descriptive values but did not conduct any significance tests. Five studies did not detect any significant differences in password strength measures. Furthermore, $N = 2$ reported a descriptive increase in users' perceived security (awareness) and two studies did not report any strength-related outcomes.

Yet, identifying factors potentially impacting password meter effectiveness warrants further analysis of significantly positive outcomes and non-significant differences taking into account the studied password meter features and study designs.

Positive Outcomes:

Of the $N = 19$ studies that detected a significant positive outcome, the majority ($N = 18$) used a between-subjects design. A total of $N = 14$ studies analysed artificial accounts, $N = 5$ studied passwords for actual accounts. The samples were recruited from: university settings $N = 8$, crowd-sourcing services $N = 7$, websites/apps $N = 3$, and other $N = 1$. This indicates that password meters can be effective in artificial as

well as actual account password creation settings, and across different groups of users.

Nineteen studies analysed 62 password meter variants. We compared which types of feedback nudges were used and how many of the variants were found to be effective over some kind of control (indicated by the number in brackets) [4]: colored feedback bar $N = 31$ (26), textual feedback $N = 11$ (7), social comparison $N = 8$ (7), fear appeal $N = 4$ (3), motivation/gamification $N = 3$ (3), affect/emotion $N = 2$ (1), password lifetime $N = 1$ (1), feedback across accounts $N = 1$ (1), and decoy effect $N = 1$ (0). This can be interpreted in that different types of framing or strength visualizations, i.e. different types of feedback nudges, have the potential to impact password strength positively. Furthermore, one might conclude that the design of the password meter did not matter a lot, e.g. as found by Ur *et al.* [111] with regards to the design of the feedback bar. Yet, in 13 of the 16 instances in which password meter variants were compared to other password meters, e.g. some kind of baseline meter including only a colored bar and/or textual feedback, significant differences were detected. For example, Egelman *et al.* [27] found longer passwords, and Dupuis and Kahn [24] higher scores, using a social comparison as compared to a colored feedback bar. In Gulenko's [41] study a password meter invoking positive emotions resulted in more words in passphrases compared to negative emotions. A gamified approach by Ophoff and Dietz [71] resulted in less predictable passwords, as compared to a colored feedback bar. Also, linking password strength to expiry, coupled with a reminder and a nudge-like visualization, produced stronger passwords, as compared to other previously tested visual password meters and nudges [84]. This indicates that the design of the password meter and type of included nudge is indeed of relevance. However, due to the various differences in the studied meters and designs, the difference cannot be traced back to certain element and requires further research.

Finally, 21 of the effective password meter variants given in brackets above included further (dynamic) password guidance on how to create strong passwords, indicating that this might be a relevant factor that can support users in creating stronger passwords, also advocated by [111, 109, 42, 54, 116].

No Differences:

Why were some meters ineffective if other studies determined efficacy? Analysing the insignificant findings in $N = 5$ studies, we were especially interested in possible explanations for these outcome.

Of the $N = 5$ studies that did not detect any significant differences, all used a between-subjects design. This might not be significant given that most of the studies overall ($N = 29$) used this design. The 22 password meter variants were of varying kinds (feedback bar $N = 4$, motivation/gamification $N = 6$, social comparison $N = 4$, textual feedback $N = 8$). Three studies used crowd-sourcing services, and two studied university samples.

[4]Please note that in some cases the control included other experimental conditions.

As no other noticeable differences or commonalities were revealed, we considered the studies more closely. One was a pilot study with a sample size the authors considered too small to reveal significant differences [25].

Egelman *et al.* [27] found significant differences in a laboratory study with actual accounts but not in the succeeding online study with artificial accounts, attributing the insignificance to people perceiving the account to be unimportant.

Segreti *et al.'s* study [89] were testing the value of password guidance, when combined with adaptive password policies and blacklists. Their findings are likely to have been impacted by the presence of these extra factors.

Renaud *et al.* [80] studied the impact of different password nudges on password creation in the wild. They advance a number of potential reasons for their insignificant findings, including a lack of password information, password reuse, and decreased variance of the deployed password strength score metric.

Golla *et al.* [38] compared four password meter visualizations to a control with no meter. All conditions included a link to additional password information but only 4% clicked the password information across all conditions. Yet, a difference to other studies was the setting. The authors recruited participants in the university foyer with a cover story, asking them to create a password for a new university portal at a stall set up in the foyer. This is an unrealistic and unlikely setting for a password creation task, with many people likely to have been milling around.

Further Interesting Findings

In addition to the security measures, many studies included usability measures such as password creation time or memorability/recall, and measured user perceptions of the password meters. Not all of these measures are listed in detail because the analysis focused on password strength.

In a total of 17 studies, memorability was analyzed. Some studies, especially online and laboratory studies, used recall rates as a proxy for memorability. Others, especially field studies, measured recall rates via the number of forgotten or reset passwords. The findings show that passwords created in the presence of password meters were not associated with lower recall rates in 14 studies. This leads to the conclusion that passwords created in the presence of password meters, even though often longer and stronger, are not necessarily less memorable. Only in three studies, including Sun *et al.'s* [104] Android Pattern Unlock study and the two instances in which password strength was coupled with password expiry [8, 84] in university accounts, did a higher number of resets occur (either due to expiry or forgotten passwords).

Some studies found that it took people longer to create a password with a password meter. Suggested reasons are longer texts and instructions which take time to read [57], increased exploration of the tool [56], and more password editing, e.g. [111].

Comparisons with password policies, e.g. [57, 59, 109], provided an indication that passwords created with a password

meter encouraged stronger passwords than the exclusive use of a password policy, even when the password meter did not enforce minimum strength requirements [125]. An explanation might be that password policy requirements constituted a target rather than a minimum, as suggested by Shay *et al.* [93]. Future work might thus systematically compare the impact of policy enforcement, optional password meter guidance, and the combination of minimum requirements with password meter guidance as successfully studied by Ur *et al.* [109].

Two studies did not apply password meters to classical text passwords, but to the Android Pattern Unlock, a graphical authentication scheme. The successful application [98, 104] serves as an indicator that feedback meters are likely to be helpful beyond the narrow context of textual password creation.

WHAT MAKES EFFECTIVE PASSWORD METERS?

Due to the exploratory nature of the analysis, some ambiguities in the classification process, and the multiple differences between study methodologies and password meters, no definite conclusions could be drawn. Even so, the results of the literature review reveal some interesting directions for future research.

First, the analysis revealed that password meters were effective more often than not across different study contexts and participant groups, while not degrading memorability. This suggests that password meters are indeed a promising strategy for encouraging more secure password creation.

Second, effective password meters included various types of feedback nudges revealing the potential for nudging users towards secure passwords without enforcing strict rules. Furthermore, this indicates that suitable visualisations of the password feedback may be helpful to the user. In line with that, Ur *et al.* [111] concluded that the combination of text and a visual indicator seems to be an important influence during password creation. While different types of feedback nudges were found to be effective, significant differences between password meter variants in some studies suggest that the password meter and feedback nudge design may well be significant and require further systematic comparison.

Third, many of the effective password meter variants included additional password information. Thus supports the assumption that their inclusion may be beneficial, not only in terms of password strength, but perhaps also in supporting the user's understanding of what makes strong passwords. In line with that [42] (p.9) stated that "*the presence of text feedback advising participants on how to make their password stronger led to stronger, more complex passwords across all participant groups*".

In summary, the results of the literature research suggest that effective password meters often combine the following elements:

- *Feedback on current password strength* in textual or visual form, i.e. the current status.
- *A feedback nudge* encouraging people to increase password strength (e.g. a social comparison nudge or a fear appeal).
- *Information on what makes a strong password*, or how to increase password strength, i.e. the desired state and how to get there.

RESEARCH QUESTIONS AND HYPOTHESES

Based on the literature review, it we can conclude that the combination of password information (current and desired state) and an intervention motivating people to adapt their current password creation routines towards the desired state, is more effective in increasing password strength than either on its own. This combination is termed a "hybrid password meter".

This leads us to the following hypotheses:

- H_1**:** A password meter combining information and a feedback nudge leads to increased *password strength* values than either on its own.
- H_2**:** A password meter combining information and a feedback nudge leads to *longer passwords* than either on its own.
- H_3**:** A password meter combining information and a feedback nudge leads to *higher password entropy* than either on its own.

Yet, even if a hybrid password meter is provided, it remains unclear how the design of the feedback nudge influences password strength. This is particularly so, given that a number of previous studies revealed significant differences between different password meter variants. It is true that different password meter feedback nudges have indeed been studied. However, it is difficult to draw comparisons because of the varying methodologies, samples, and levels of password information provided. This research thus aims systematically to compare different feedback nudges in hybrid meters that differ only in the design of the included feedback nudge.

Aside from analysing the impact of different password meters based on different feedback nudges on password strength, we were also interested in password memorability, and in the users' perceptions of the password meter. This led us to the following research questions:

How do password meters based on different feedback nudges, e.g. targeting different biases, heuristics and norms, impact:

RQ_1**:** password creation, in terms of password strength, length and entropy?

RQ_2**:** password memorability?

RQ_3**:** users' perceptions of the password meter and the password creation process?

PASSWORD METERS

In the following sections, we describe the password meters evaluated in our study. Overall, the password meters can be classified as targeting three main areas as depicted in Figure 4: the person, the password creation context, and the social context. The following sections describe first the original password meter by Ur *et al.* [109], and then the variations analysed in our study.

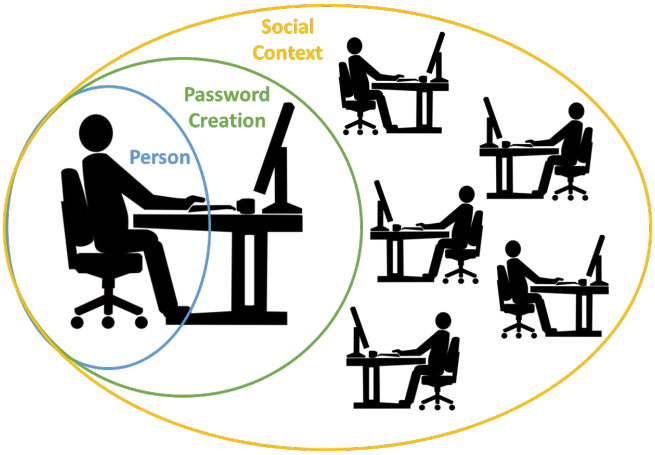

Figure 4. Graphical depiction of the areas targeted by the different hybrid password meters. Image adapted from www.pixabay.com.

Original/Control

As a basis for developing multiple variations and an already successfully evaluated control condition, we used the password meter designed by Ur *et al.* [109]. The developed password meters were compared to an existing meter, rather than nothing at all, as suggested by Heidt and Aviv [48].

As shown in Figure 5, the original hybrid password meter provides users with a colored bar to indicate password strength, and information on how to increase password strength that dynamically changes with the user's input. The link "How to make strong passwords" provides general information on how to create strong passwords. The strength estimation in this password meter is based on 21 heuristics and a neural network as described in [109]. The resulting password strength can be converted to a score ranging from 1 to 100. As the password meter combines information provision with as simple color-coding nudge it is labelled a "Hybrid Password Meter".

We chose this hybrid password meter as a basis for the other variations for the following reasons:

- It has been shown to encourage stronger password choices, as compared to plain password entry fields.
- The scoring and the type of textual feedback are based on a line of research including state-of-the-art knowledge, in terms of password creation and meters. As such it can be viewed as an "umbrella" password meter that includes many of the insights gained from related work.
- Furthermore, the design of the password meter is close to real-world examples enumerated by [111]. This includes password meters using feedback bars and/or textual information including Google, Yahoo, Twitter or Paypal.
- The password meter includes potentially effective elements identified in our literature review that we aim to study in more detail: an indication of password strength in the form of a feedback bar (current status), information about what makes a strong password and instructions on how to improve it given the user's input (the aim and how to get there), and a nudge (color-coding of the bar).
- The tool and code are open source and easy to adapt to the purposes of our study and include different nudges. This makes the design and scoring transparent for other researchers who aim to conduct similar studies.

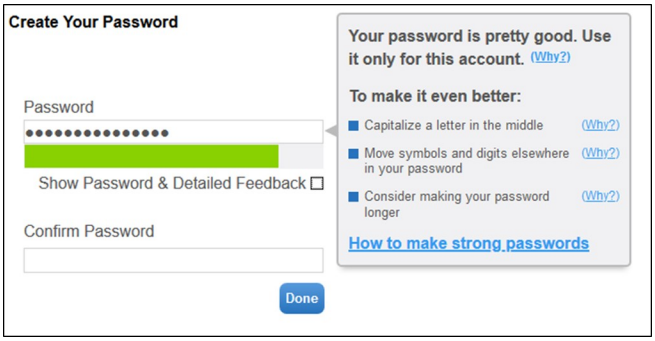

Figure 5. Graphical depiction of the original hybrid password meter based on Ur *et al.* [109].

To test the single versus combined impact of the interventions, we also included two variations of the original hybrid password meter as control conditions. These variations are: (1) a simple password meter only showing the colored bar (Figure 6), and (2) a simple password meter only containing the information (Figure 7). Variation 1 is labelled "Simple Nudge" as only the color-coding nudge is provided, and variation 2 is labelled "Information".

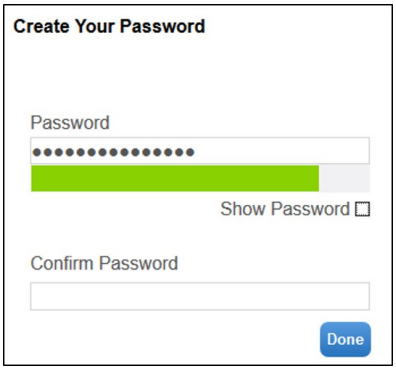

Figure 6. Graphical depiction of the simple password meter in the control condition containing only a visual nudge, in the form of a feedback bar.

The Person

Figure 8 provides an overview of the two hybrid password meters targeting the person by using a (positive) fear appeal nudge and a motivation nudge.

(Positive) Fear Appeal Nudge

According to Witte [124], for a fear appeal to trigger protection motivation, it is important for perceived threat *as well as* perceived efficacy to be high. A fear appeal should thus include elements of both: (a) the severity of, and susceptibility

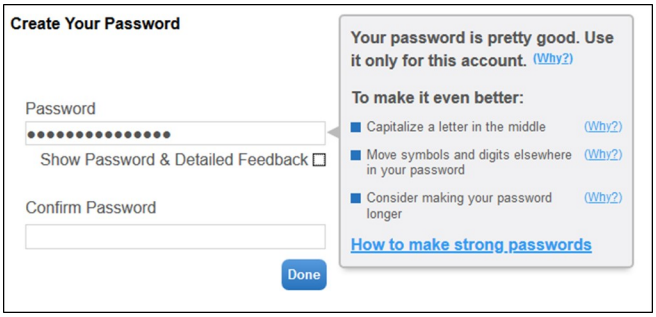

Figure 7. Graphical depiction of the simple password meter in the control condition containing only password information.

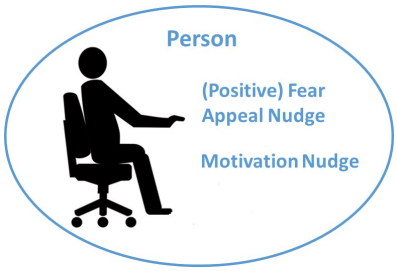

Figure 8. Graphical depiction of the two hybrid password meters targeting the person. Image adapted from www.pixabay.com.

Figure 9. Graphical depiction of the hybrid password meter based on a (positive) fear appeal nudge. Image adapted from www.pixabay.com.

to, the threat, and (b) the efficacy of the response and the individual to take action. A mismatch, such as a high perceived threat but low perceived efficacy can lead to maladaptive behaviours, such as denial. A fear appeal should thus not leave people afraid but help them cope with the perceived threat by providing useful suggestions.

Along these lines, a recent review by Renaud and Dupuis [78] on the use of fear appeals in the cybersecurity domain revealed mixed findings in terms of their effectiveness, perhaps due to different levels of fear induced, different types of measures, and different contexts of application. The review highlights the importance of further experimental studies to provide "*a more clear-cut judgement about the utility of fear appeals in the cyber security domain*" [78] (p.10). The authors provide guidelines for subsequent studies including the use of a randomized controlled design, due consideration of ethical issues, and the evaluation of post-appeal measures of perceived fear, self-efficacy, severity and susceptibility.

Following the findings of Renaud and Dupuis [78], and to comply with ethical considerations to avoid inducing negative feelings in participants [81], we decided to frame the fear appeal nudge in a positive way. That is, instead of aiming to create fear by showing how quickly an attacker would crack the provided password, we aimed to increase protection motivation by framing the message in terms of how much the time to crack a password increases, and how that makes attacking the password increasingly unattractive to a potential attacker. This was visualized by using an image of an attacker aiming to crack the password (see Figure 9). Upon entering a weak password, the attacker is smiling and says that the password was easy to guess. With improved password strength, the attacker becomes increasingly frustrated and states that the password was very hard to guess. The word "password" in the image is increasingly replaced with question-marks to highlight this fact. Similar to the other conditions, the color changes from red for weak progressing to yellow and then to green for strong passwords.

Furthermore, pre- and post-appeal measures of the users' perceptions as described above were implemented. The pre- and post-measures of fear and self-assurance were based on the related PANAS-X scales [119], a well-established instrument for measuring affect. Items for measuring perceived threat, perceived susceptibility and perceived severity were based on questionnaires developed within the framework of the Technology Threat Avoidance model proposed by Liang and Xue [60] and variations of these items as described by Arachchilage and Love [6] as well as Gerber *et al.* [35].

To provide people with an indication of how long their password would resist an attack, we relied on the `zxcvbn` password meter [122] estimator. This also inspired the password meter score calculated by Ur *et al.* [109] and thus provides a consistent measure. Moreover, it has been found to be a relatively accurate estimate in terms of actual resistance to online attacks [122]. We displayed the time a password would resist a throttled offline attack assuming 10,000 attempts per second. To derive the equivalent for different password scores we created a number of passwords with different scores on the password meter developed by Ur *et al.* [109] and mapped those to the times estimated by the `zxcvbn` meter for the same passwords. For example, a password with a score between 6 and 10 equalled a time estimate of two seconds, while passwords with a

score between 61 and 65 would resist guessing attacks for about 12 days.

Motivation Nudge

The motivation nudge aimed to increase the users' motivation without eliciting fear. Referring back to the description of the fear appeal nudge, the motivation nudge targeted the self-efficacy and efficacy of the response. The instructions thus stated that creating strong passwords was challenging but assured the user that he/she would be able to do it by following the provided suggestions. The visualization, depicted in Figure 10, showed a little runner that was positioned at a starting line when the user started creating a password and then began to run towards the finishing line as password strength improved. The finishing line was crossed when the password score reached 100. The final image was that of a happy runner jumping around with a little gold medal around his neck. In addition, as the password score increased, encouraging statements were displayed such as "You are nearly there" or "Take a final leap".

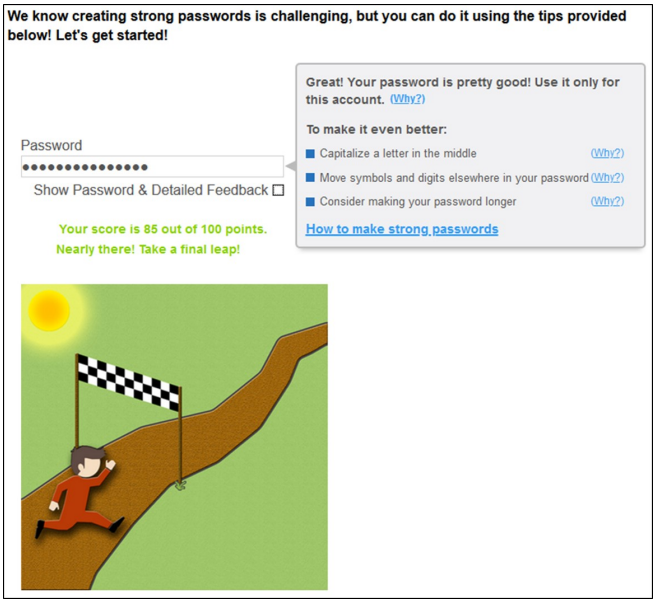

Figure 10. Graphical depiction of the hybrid password meter based on a nudge targeting self-efficacy.

The Password Creation Context

Similar to other security tasks, authentication, and in this case password creation, is not an objective in itself but a secondary task. This means that people do not authenticate to authenticate, but rather to gain access to accounts and data. They might need to send an email, check their online banking account balance, or access work documents. Hence password creation and the password login process represent a hurdle users have to overcome. Furthermore, from a user perspective, stronger passwords require more effort, i.e. increased length and complexity often decrease memorability and increase password entry times. The two hybrid password meters described below, and named in Figure 11, aim to address and mitigate that perceived increase in effort, and to balance the advantages and disadvantages of weak vs. strong passwords to facilitate strong password choice.

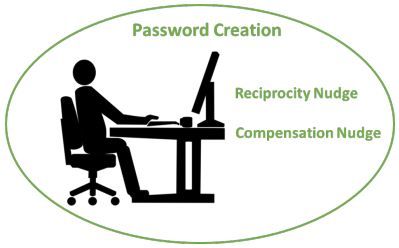

Figure 11. Graphical depiction of the two hybrid password meters targeting the password creation context. Image adapted from www.pixabay.com.

Compensation Nudge

The first hybrid password meter using a compensation nudge, aims to compensate the users' increased effort involved in choosing stronger passwords by extending the password's lifetime, that is, the time a password is valid before it expires.

Regular password expiry, regardless of password strength, has long been common practice in industry, yet this practice has found to negatively impact password strength in the long term [17, 127]. With regular password expiry, users tend to use short and only slightly altered versions of the previous password (e.g. changing a number at the end of the password) [117, 127] to cope with the cognitive effort required to memorize multiple passwords.

What is different with our compensation intervention is that increased password strength is compensated by extending the password's lifetime depending on the score calculated in the password meter. Very strong passwords, i.e. passwords with the highest score of 100 points, do not need to be changed at all unless there is evidence of compromise, as suggested by organizations such as the National Institute of Standards and Technology (NIST) [39]. This allows users to choose between the more balanced options of having to change a short, easy-to-memorize password frequently, or to memorize and type a potentially longer password that does not need to be changed. To nudge users towards the second option, an image of a calendar as a metaphor for lifetime was displayed. With increasing password strength, the cross marking the password change date moved to a later date (and vanished when a score of 100 was reached) while the color the lifetime was shown in changed from red to yellow to green (see 12). While linking password strength to lifetime has been successfully tested by [84] in a longitudinal field study, it produced mixed findings in a study conducted by Becker, Parkin, and Sasse [8]. They found an increase in password strength but also that very strong passwords led to high reset rates.

Reciprocity Nudge

The hybrid password meter using a reciprocity nudge was based on the principle of responding to a received, positive action with another positive action, e.g. by giving something back. In this case, the effort associated with stronger passwords was "compensated" by highlighting the efforts already undertaken for the sake of the user's security by the service provider. That

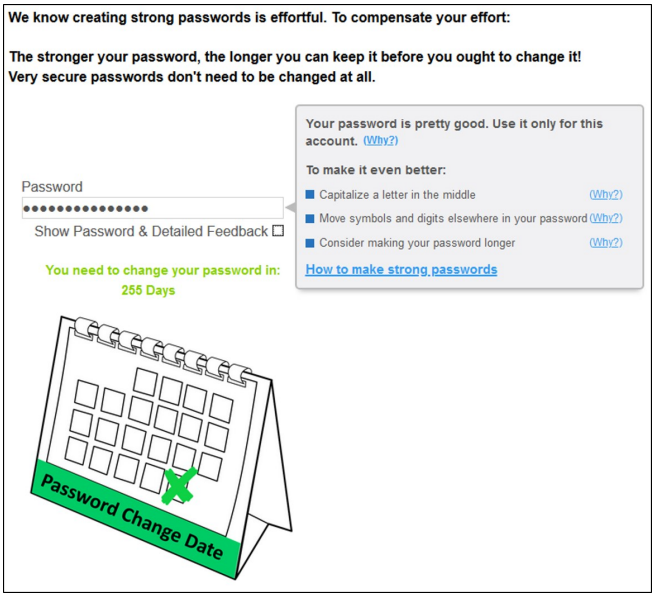

Figure 12. Graphical depiction of the hybrid password meter based on a nudge compensating effort with increased password lifetime. Image adapted from www.pixabay.com.

is, explanations provided as to the security efforts implemented by the provider, such as hashing and salting of passwords and securing the transfer of the password to the server. Within the interface, the password meter was split: on the left-hand side the participant saw a list of the implemented security measures, along with a link that provided an explanation. The left-hand colored bar labelled "Technical Security" was colored green, indicating that sufficient measures had been implemented. The instruction then asked the user to do their part in securing the system by following the suggestions displayed on the right-hand side and choosing a strong password. The right-hand side colored bar was empty in the beginning and filled in as password strength increases while changing color from red to green analog to the bar in the original condition. A screenshot of the reciprocity nudge is shown in Figure 13.

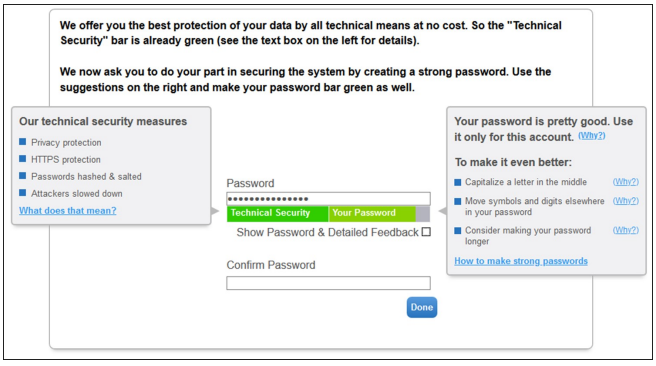

Figure 13. Graphical depiction of the hybrid password meter based on a reciprocity nudge.

The Social Context

Social norms are implicitly or explicitly communicated attitudes and rules that describe what kind of behaviour is deemed appropriate or inappropriate for a certain situation by the members of a social group [36, 105]. Social norms, and the potential negative consequences of not following them, can be observed in the interactions within social groups. Sunstein [105] compared social norms with behaviour-related "taxes" or "subventions" that can influence behaviour. In the psychological literature, a large number of and partially overlapping types of social norms are differentiated [19]. One of the well-known theoretical approaches, the Focus Theory of Normative Conduct by Cialdini, Kallgren and Reno [20], differentiates between injunctive and descriptive social norms. The injunctive norm describes the perception of what behaviour is accepted or wished for and includes the perceived pressure to follow these norms to avoid potential social sanctions, i.e. the norm describes what other members of the group want oneself to do. The descriptive norm describes the perception of the behaviour actually shown by the other members of the social group, i.e. what the majority of members actually do. Cialdini *et al.* [20] propose that norms especially influence behaviour when they are activated, i.e. made salient in a certain situation. For example, it has been shown that people tend to drop rubbish more often when the environment is already littered (descriptive norm), and that a reference that littering is not acceptable made people drop less rubbish (injunctive norm) [20]. The example also shows that to encourage people to engage in a desired behaviour, the descriptive norm on its own might be counterproductive if many people exhibit the undesired behavior. This effect can be mitigated by combining it with an injunctive norm as has been demonstrated with energy use feedback that combined an indication of energy use (descriptive norm) with a sad or happy smiley (injunctive norm) [88].

Figure 14 shows the social context and the two hybrid password meters making use of social norms.

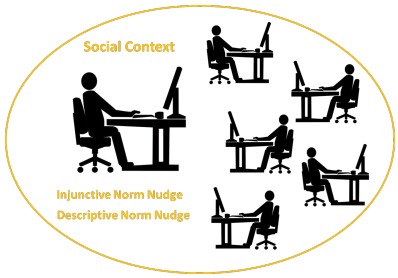

Figure 14. Graphical depiction of the two hybrid password meters targeting the social context. Image adapted from www.pixabay.com.

Descriptive Social Norm Nudge

The design of the hybrid password meter using a descriptive social norm nudge as shown in Figure 15 was based on a pilot study conducted by (blinded for review). The password meter compares the input of the user to that of a fictional "average" user. The password score calculated by Ur *et al.* [109] is used to indicate what percentage of users have a better password score. The colored bar was replaced by a visualization of

the average user that is increasingly filled and dynamically changes its color from red to green with increasing password strength. Also, the dynamic text was adapted to highlight the comparison, e.g. "Your password is pretty good" was changed to "Your password is pretty good as compared to other users". Research showed that descriptive norms alone might lead people that already behave "wisely" to reduce their behavior to the average level, such as in the case of energy-saving families increasing their consumption after becoming aware of the comparison [88]. Therefore, similar to the study described by Schultz *et al.* [88] the visualization included a sad or happy face depending on password strength to provide users with an indication of what is desirable to mitigate that potential side-effect.

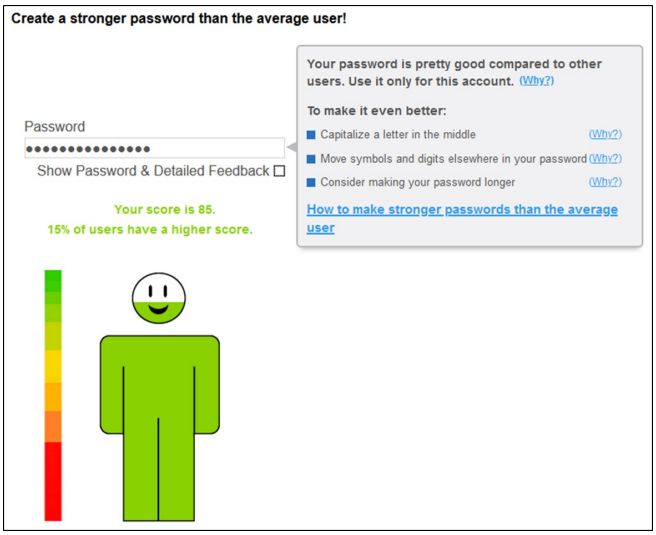

Figure 15. Graphical depiction of the hybrid password meter based on a descriptive social norm nudge.

Injunctive Social Norm Nudge

The hybrid password meter using an injuctive norm nudge aimed to provide users with a indication of what is viewed as desirable by the social group, i.e. other users of the fictional service. It was stated that other people wanted the participant to create a strong password so that their accounts would be protected. The instruction explained that this was desirable as every hacked account also poses a potential threat to other users' data. An attacker could use a hacked account to send malware-infected files to the participant's contact list, and thereby also infect their PCs, or send seemingly innocuous phishing emails. As described above, not following an injunctive norm "ideal" is often associated with some kind of social sanction. That was implemented by visualizing social disapproval with non-compliance and rewarding compliance with social approval. The feedback meter took the form of ten smileys that each represented ten percent of the user base. With increasing password strength, the percentage equivalent to the password score from 1 to 100 turned from a red disapproving face to a green happy face accompanied with a "thumbs up" symbol frequently used by social networks such as Facebook. For example, if the password score was 60, six of the ten faces were green and four were red. A score in between the tens,

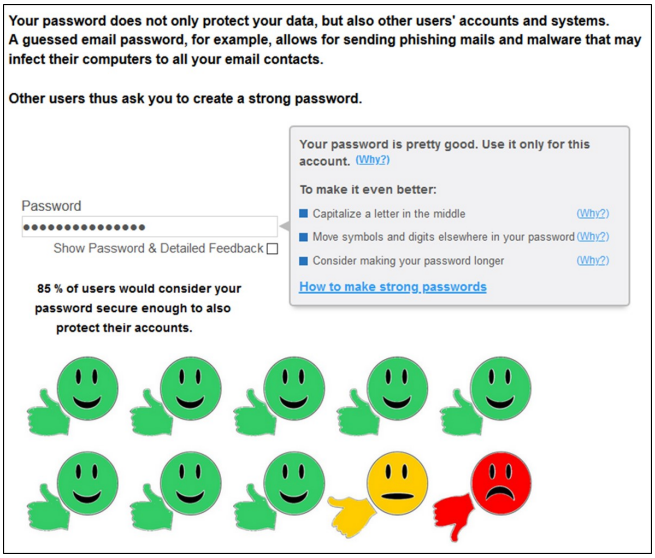

Figure 16. Graphical depiction of the hybrid password meter based on an injunctive social norm nudge. Image adapted from www.pixabay.com.

e.g. 64, was visualized by a yellow, neutral face to indicate an upward trajectory.

METHOD

To compare the different password meters that were based on different types of feedback nudges, we conducted an online study with N=645 participants. Each participant interacted with one password meter, thus the design was a between-subjects design. The strength estimation of the password meters, and the general information on password creation, was based on a password meter created by Ur *et al.* [109]. This was constant across all experimental conditions, while the way the feedback was presented or framed was adapted to target different heuristics and norms (see section *Password Meters*) to constitute the different experimental conditions.

Procedure

We conducted an online study using Amazon Mechanical Turk to evaluate the effectiveness of the different password nudges on participants' password strength scores, length and entropy. Password strength was measured as a score ranging from 1 to 100 based on the scoring described by Ur *et al.* [109] that combines several heuristics (e.g., the inclusion of common words or keyboard patterns) with a neural network to model password guessing attacks. Password entropy was measured in bits based on the Shannon entropy measure [92] that describes the level of information or uncertainty a password holds. Password length was measured as the number of characters in the password.

Following an introduction and an informed consent page, in line with strict national data protection guidelines, the participants were required to pass two attention check items that were based on the suggestions by Meade and Craig [64] to identify careless responses in survey data that have already

been applied in the cybersecurity context by Egelman and Peer [26].

On the next page, participants were required to create a password for a fictional online service. Based on a study by Ur *et al.* [109] participants were asked to role-play to create a password for an important online service they cared a lot about, such as their primary email account. They were asked to create a password not recently used, to create and handle the password as they would do with their actual passwords, and to be prepared to log in with the password again. The participants were randomly assigned to one of the password creation conditions in a between-subjects design. Participants were not allowed to complete the study using a mobile phone as previous research found this input method to be more error-prone and frustrating, and passwords to be weaker [65]. To avoid bias and problems with the visualization displays, users were asked to use a PC or laptop.

Following the password creation process, the participants were asked to rate the password meter, the password they created, and how they felt creating a password with the help of the password meter. Some items were inspired by the ones asked by Ur *et al.* [109] for comparability. However, we refined and extended the scales to rate the three aspects stated above. Furthermore, we changed the scale to a semantic differential similar to the questionnaire "AttrakDiff" [46] that is used to evaluate several usability aspects of products or services. An exemplary screenshot of the semantic differential is shown in Figure .

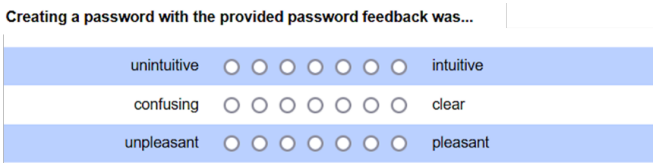

Figure 17. Exemplary screenshot of the semantic differential used to explore the participants' perceptions of the password meters.

Afterwards, the participants were asked to rate the different features of the password meter, including the text feedback, the suggested improvements, and the different visual nudges aimed at increasing password strength. To aid memorability, for each item set a screenshot of the relevant feature was provided again. Individual measures were taken to analyse whether one design element of the overall combination was considered especially useful.

As control variables, we measured the participants' affinity for technology interaction with the ATI scale by Franke *et al.* [30], their attitude towards choosing password using items of the security behavior intentions scale (SeBis) by Egelman and Peer [26] and demographic information such as gender, age, education, and occupation. Because some of the password interventions targeted social norms we also asked for the users' compliance with social norms using the six items with the highest factor loadings of the Social Norm Espousal Scale by Bizer *et al.* [12] and some self-constructed items.

Finally, we measured memorability via password recall rates a frequently used proxy in the related work studied in our literature review. To do so, we asked the participants to re-enter the password they created in the beginning of the study. They were given three opportunities to replicate their original password, and received feedback on whether the entered password was correct. After the third unsuccessful attempt, participants were forwarded to the next questions asking how they created and stored the study password, e.g. whether they had used a password manager to generate and remember a password for them. A complete list of the scales and items used in the study is provided in Appendix C.

After a retention of two weeks, all participants were asked to provide the previously-created password again, to test memorability via password recall rates in a follow-up study. Each participant was, once again, given three opportunities to provide it. After a couple of days, all participants who had not yet provided their password received a reminder.

Sample

The sample consisted of 645 English-speaking people residing in the United States aged 18 years and older. A total of 255 identified as female, 379 as male, and 2 as diverse. The remaining participants preferred not to say or did not provide an answer. A total of 229 people had some kind of IT (security) background, that is, studied computer science, IT security, or were in an occupation in that area of expertise. Table 2 provides more information about the sample's age distribution, education, and occupation.

Ethical Considerations

The study was planned and conducted in line with the ethics checklist provided by our university's ethics committee and guidelines for ethical psychological research [69]. All participants were presented with an informed consent sheet containing information about the aim of the study, the procedure and the expected duration. Further, participants received information about the handling of their data in accordance with strict data protection laws, their rights in terms of data privacy and the contact information of the researchers to answer questions or concerns. The survey was implemented in an online survey tool (blinded for review) that stores data in accordance with strict data protection laws. The collection of demographic information was reduced to a minimum of relevant control variables: gender, age, education and occupation. Age was collected in age ranges to enhance anonymity. The participants were compensated based on a $10/hour rate, which exceeds the country's minimum wage.

RESULTS

We present the results structured by our research questions.

RQ_1: Password Strength, Length and Entropy

When asked for the applied strategies for creating the password within the study, the majority of participants ($N = 521$) said that they created a new password by themselves. Those participants who used a password manager ($N = 69$) or those who reused an existing password ($N = 57$) were excluded from

Measure	N	%
Age (in years)		
18-24	41	6.4
25-34	315	48.8
35-44	172	26.7
45-54	67	10.4
55-64	31	4.8
>64	10	1.6
No answer	9	1.4
Education		
Finished High School	141	21.9
Associate Degree	88	13.6
Bachelor Degree	323	50.1
Master Degree	69	10.7
PhD or similar	10	1.6
Other/ No answer	14	2.2
Occupation*		
Pupil/ In School	2	0.3
College/ University student	37	5.7
Employee / Civil Servant	460	71.3
Civil Servant	10	1.6
Self-Employed	115	17.8
Unemployed/ Seeking Employment	30	4.7
Retired	7	1.1
Other/No answer	12	1.9

Table 2. Description of the sample in terms of age, education, and occupation. *Multiple answers were possible.

the password strength, length and entropy analysis. The reasoning was that in these cases no password creation process took place, and the process could therefore not have been influenced by the meter. Furthermore, an inclusion of reused passwords, that were rather weak with a password strength of $M = 40.21$, $SD = 30.91$, $Md = 35.05$, or system-generated passwords, that were rather strong with $M = 89.99$, $SD = 16.20$, $Md = 95.38$, might have biased the comparison between groups.

As neither password strength nor entropy were metric measures, and password length was not normally distributed, nonparametric and independent-samples Kruskal-Wallis tests were conducted to check for overall differences on a significance level of $p = .05$. A table providing an overview of the descriptive strength, length and entropy values is provided in Appendix D.

Selective follow-up comparisons to test:

(a) the combined vs. single impact of information and a feedback nudge, and

(b) differences between the control condition and the experimental conditions

(c) differences between nudges targeting the person, the password creation context, and the social context

were carried out using Mann-Whitney-U tests. To account for multiple tests, the Benjamini-Hochberg correction [11] was applied to all follow-up comparisons. The corrected significance level was $p = .022$ for password strength and entropy, and $p = .026$ for password length.

Comparing the passwords across all nine conditions revealed that there were significant differences in terms of password strength ($H(8) = 31.72$, $p < .001$), length ($H(8) = 29.56$, $p < .001$) and entropy ($H(8) = 29.75$, $p < .001$).

The combined vs. single impact of information and feedback nudge

To test H_1, H_2 and H_3, first the combined *vs.* single impact of interventions within the control conditions was considered before comparing them to the experimental conditions. That is, the variations of the original hybrid nudge password meter.

Original/Control Conditions

A visual inspection of the boxplot shown in Figure 18 shows that the password strength medians are lowest for the control conditions information (only showing password information) and simple nudge (only including the colored bar) followed by the original hybrid password meter by Ur *et al.* [109].

One-sided follow-up tests reveal that the strength differences between the original hybrid password meter and the information or simple nudge condition are not significant ($p = .112$ and $p = .043$ respectively). The differences in terms of length and entropy are significant when the original hybrid password meter is compared to the information condition ($Z(1) = -2.05$, $p = .020$, $r = .19$ and $Z(1) = -2.08$, $p = .019$, $r = .19$), but not when compared to the simple nudge ($p = .040$ and $p = .052$ respectively).

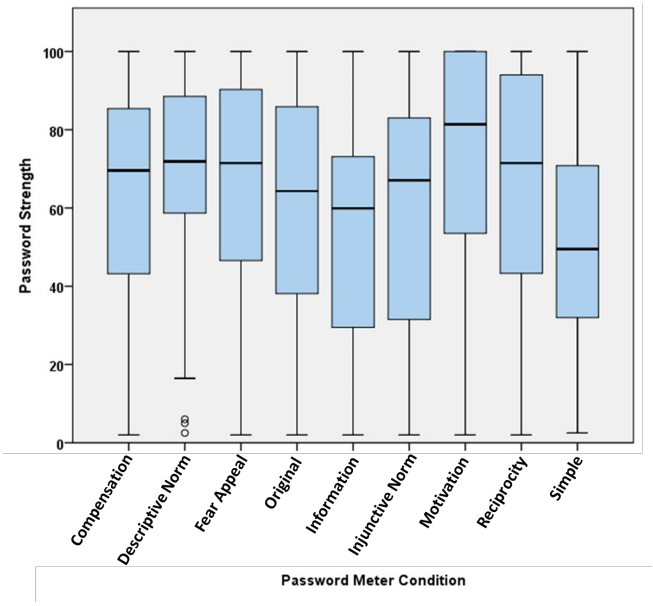

Figure 18. Boxplots comparing password strength across conditions.

Comparison of control and experimental conditions

The simple nudge, information and original hybrid password meter condition were compared to the six hybrid password

meter variations in one-sided follow-up comparisons to test H_1, H_2 and H_3.

Table 3 shows that the password strength, length and entropy values in the feedback nudge variations were significantly higher than those in the information or simple nudge condition in most cases. However, compared to the original hybrid password meter the strength, length and entropy values were not significantly higher, except for the motivation nudge. All p-values that remain significant on the corrected significance level are marked with an asterisk. The effect size r can be interpreted as follows: values smaller than .3 can be considered a small effect, values between .3 and .5 medium effects and values above .5 large effects.

The Person

On a scale from 1 to 7 the N = 73 participants in the fear appeal condition rated their susceptibility to an attacker trying to guess their passwords as M = 3.74 (SD = 1.93, Md = 4). The threat posed by such an attempt was rated as M = 4.60 (SD = 1.79, Md = 5). The severity of a password guessing attempt (M = 4.44 (SD = 1.88, Md = 5) was rated significantly lower than the severity of a successful password guess by an attacker (M = 5.23 (SD = 1.66, Md = 6) with $t(71) = -4.06$, $p < .001$.

The fear appeal nudge did not significantly impact the fear levels that were M = 1.88 (SD = 1.16) before and M = 1.93 (SD = 1.16) after password creation with $t(72) = -.877$, p = .383, d = .04. Likewise, the self-assurance levels before (M = 2.87, SD = 1.05) and after (M = 2.99, SD = 1.08) measured with the related sub scales of the PANAS-X [119] on a scale ranging from 1 to 5 did not change significantly with $t(72)$ = -1.69, p = .095, d = .11.

Even though the descriptive password strength, length, and entropy values of the motivation nudge were higher than that of the (positive) fear appeal nudge, the differences were not significant on the corrected significance level (p = .039, p = .034, and p = .035 respectively).

The Password Creation Context

Comparing the two nudges targeting the password creation context, the compensation and the reciprocity nudge, revealed no significant difference in terms of strength (p = .395), length (p = .367), and entropy (p = .357).

The Social Context

The complete sample rated the six highest loading items of the social norm espousal scale with a mean of M=3.30 (SD=.95, Md=3.5). The self-created social norm items were answered with a mean of M=3.03 (SD=.86, Md=3). A comparison between all nine conditions revealed that there was no difference, neither for the social norm espousal scale ($F(8, 630) = .520$, p = .842, partial η^2 = .007) nor for the self-constructed items ($F(8, 630) = .980$, p = .450, partial η^2 = .012). Thus, the participants that were shown a social norm password nudge before answering the items did not answer the social norm items significantly differently than other groups, and their commitment to social norms did not seem to be higher.

A comparison of the two social norm nudges, the descriptive norm nudge and the injunctive norm nudge, showed that the higher descriptive values of the descriptive norm nudge did not significantly differ from those of the injunctive norm nudge in terms of strength (p = .115), length (p = .255) and entropy (p = .461).

RQ_2: Password Memorability

At the end of the study of the 645 participants, 609 or 94.42% were able to remember or otherwise retrieve the password created in the first phase of the study. A total of 27 did not remember the password correctly, and 9 people did not provide an answer.

Of the 267 people that took part in the follow-up study that started about two weeks after the first study, 93 (34.83%) were able to reproduce the password they created (79 with the first guess, 9 with the second and 5 with the third guess). Most people reproduced their passwords in the original hybrid password meter condition (50%), the reciprocity nudge condition (46.15%), and the compensation nudge condition (44.44%). The lowest percentage of correct passwords was in the control groups that were only shown information (21.21 %) or only a colored bar (21.43%). The remaining conditions received the following memorability rates: injunctive norm nudge 40.91%, descriptive norm nudge 34.38%, fear appeal nudge 33.33%, and motivation nudge 27.78 %.

Even if we control for all participants who admitted that they used some tool to record their password, the picture only changes slightly. The hybrid password meters still cover the first six places in terms of recall rates. The simple nudge and the information conditions are in positions seven and eight in the list. The memorability in the hybrid motivation nudge condition decreased slightly from the seventh to the last place. In that context, it should also be noted that a total of 160 people in the follow-up study who admitted to having recorded their password were not able to reproduce their password correctly.

RQ_3: User Perceptions

First, the participants were asked to provide ratings on a semantic differential indicating how they perceived the password creation feedback (see Figure 19), the password they created (see Figure 20), and how they felt creating a password with the feedback (21). The profile diagrams in the Figures 19, 20 and 21 show qualitative profile diagrams for each group separated by the original/control conditions and the interventions targeting the person, the password creation context, and the social context for an improved legibility. The semantic differential scale consisted of seven points. As most means were located between the points 4 and 6, we decided to highlight that segment for increased legibility and comparability. Please note that the lines in the profile diagrams are not supposed to indicate continuous measurements but to make it easy to spot differences between the qualitative profiles and the independently evaluated adjectives.

Taking a closer look at the profile diagrams the conditions with the highest and lowest descriptive values were identified for each pair of words. Within the ratings of the password creation feedback the compensation nudge condition received the highest overall mean with M = 5.21 (SD = 1.45), and was perceived as easiest, most intuitive, clear and helpful (see the

Experimental Condition	PW Strength			PW Length			PW Entropy		
	Z	*p*	*r*	*Z*	*p*	*r*	*Z*	*p*	*r*
Comparison with Simple Nudge									
Motivation	-4.22	<.001*	.39	-3.71	<.001*	.35	-3.76	<.001*	.35
Fear Appeal	-2.92	.002*	.27	-2.23	.013*	.21	-2.32	.010*	.21
Compensation	-2.05	.020*	.19	-1.95	.052	.18	-1.60	.055	.15
Reciprocity	-2.84	.002*	.26	-2.54	.006*	.24	-2.62	.004*	.24
Descriptive Norm	-3.47	<.001*	.32	-3.12	<.001*	.29	-2.92	.002*	.27
Injunctive Norm	-1.33	.092	.12	-1.56	.060	.14	-1.70	.045	.16
Comparison with Information									
Motivation	-3.81	<.001*	.35	.3.94	<.001*	.36	-4.08	.001*	.38
Fear Appeal	-2.40	.008*	.22	-2.62	.004*	.24	-2.87	.002*	.26
Compensation	-1.59	.060	.15	-2.11	.018*	.19	-1.80	.071	.17
Reciprocity	-2.42	.008*	.22	-3.00	.001*	.28	-3.09	.001*	.28
Descriptive Norm	-2.95	.002*	.27	-3.65	<.001*	.33	-3.55	<.001*	.33
Injunctive Norm	-0.95	.171	.09	-1.66	.049	.15	-1.91	.028	.17
Comparison with Original Hybrid Password Meter									
Motivation	-3.03	.001*	.29	-2.43	.015*	.23	-2.54	.006*	.24
Fear Appeal	-1.10	.272	.10	-0.67	.250	.06	-0.83	.204	.08
Compensation	-0.44	.331	.04	-0.16	.435	.02	-0.06	.475	.01
Reciprocity	-1.42	.080	.13	-1.08	.141	.10	-1.09	.137	.10
Descriptive Norm	-1.54	.062	.14	-1.41	.079	.13	-1.28	.100	.12
Injunctive Norm	-0.19	.427	.02	<.001	.500	.00	-0.03	.372	.00

Table 3. Results of pairwise comparisons between control and experimental conditions. Z = standardized test value, p = Significance value, r = Effect size, * significant p-values after correction for multiple comparisons.

third profile diagram in Figure 19). The fear appeal nudge displayed in the second diagram in Figure 19 was perceived most fun, pleasant, and motivating whereas the motivation nudge was rated as most novel, challenging and informative. The lowest overall value was received by the information condition with $M = 4.79$ ($SD = 1.73$). It was perceived the least pleasant, clear, and motivating. The simple nudge was perceived least novel, challenging and informative, while the original hybrid password meter was perceived least fun, easy, and intuitive. This can be seen in the first diagram of Figure 19).

In line with the descriptive password strength values, the created password was rated "best" overall in the motivation nudge condition ($M = 5.46$, $SD = 1.41$) and "worst" in the simple nudge condition ($M = 5.14$, $SD = 1.41$) as shown in Figure 20. The motivation nudge password was rated most strong and secure, the fear appeal nudge most creative and stronger compared to the own passwords, and the compensation nudge password most random and better as compared to other people's passwords. The simple nudge shown in the first diagram of Figure 20 received six of the lowest ratings and was perceived the least strong, complex, long, secure, creative, and random.

The diagrams in Figure 21 shows how the participants felt when creating their password. The participants felt "best" with the fear appeal nudge ($M = 5.47$, $SD = 1.46$), especially competent, good, excited, appreciated, and proud. Other high scores were received by the compensation nudge: Participants felt equally appreciated as with the fear appeal nudge, and most capable, certain, assured, and protected. The participants felt least good with the reciprocity nudge ($M = 4.93$, $SD = 1.66$) that received the lowest scores for the words capable, certain, competent, excited, assured, and proud.

Second, the participants were asked to rate the overall and individual features of the provided password feedback. The items measuring the individual features concerned the visual nudge elements, the textual password information, and the suggested improvements for the password. On a 7-point scale from "strongly disagree" to "strongly agree" the participants "slightly agreed" that the overall feedback helped them to create a strong password ($M = 5.09$, $SD = 1.82$, $Md = 6$). They perceived the strength meter score to be accurate ($M = 5.61$, $SD = 1.48$, $Md = 6$) and agreed to have learned something new from the feedback ($M = 5.67$, $SD = 1.36$, $Md = 6$). The text feedback and the suggestions that all groups except for the simple nudge condition were provided with were rated as informative ($M = 5.77$, $SD = 1.32$, $Md = 6$ and $M = 5.76$, $SD = 1.34$, $Md = 6$). Most participants furthermore agreed that the text feedback and suggestions encouraged them to create a different password than they would have otherwise ($M = 5.24$, $SD = 1.76$, $Md = 6$ and $M = 5.30$, $SD = 1.81$, $Md = 6$).

The colored visualisations of the different feedback nudges that differed between groups (except the information condition) were rated as rather helpful ($M = 5.40$, $SD = 1.58$, $Md = 6$, informative ($M = 5.60$, $SD = 1.45$, $Md = 6$), and encouraging different-than-usual passwords ($M = 5.13$, $SD = 1.87$, $Md = 6$).

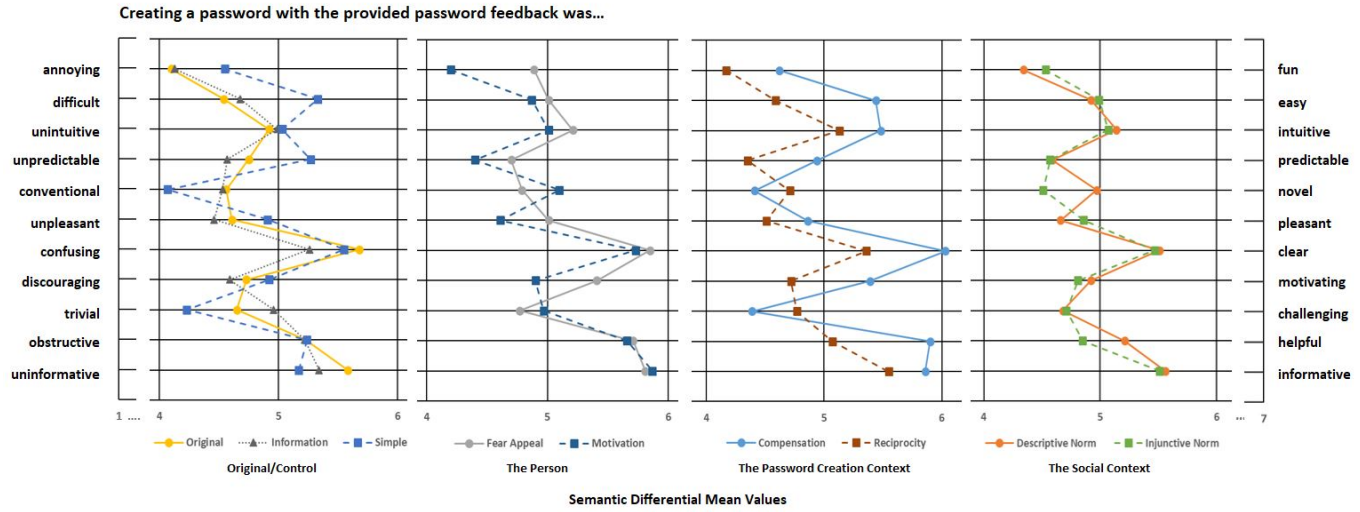

Figure 19. Profile diagrams of the user ratings' mean values in terms of the password creation feedback. Please note: The connecting lines between the items do not indicate a continuous measurement. They are supposed to increase legibility and spotting differences between qualitative profiles.

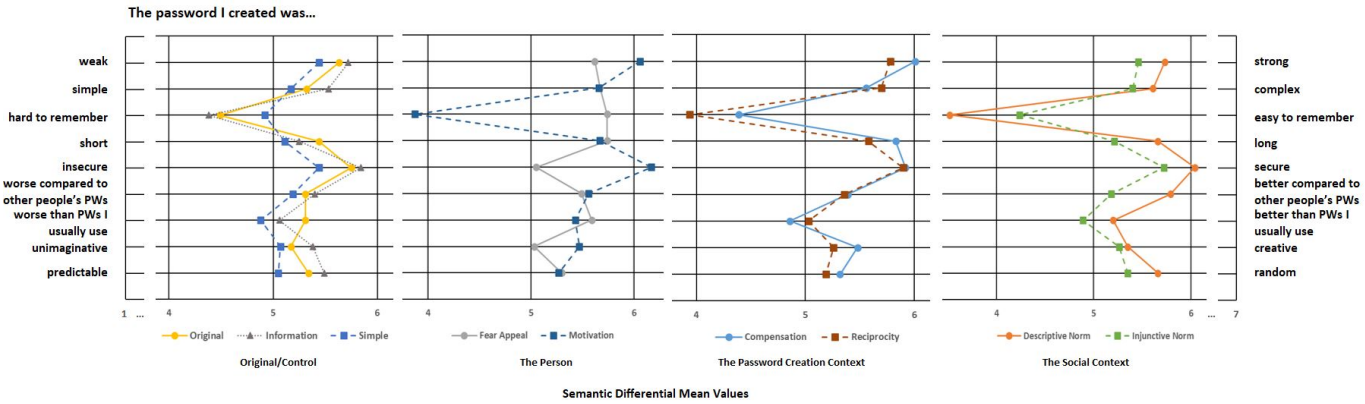

Figure 20. Profile diagrams of the user ratings' mean values in terms of the password they created. Please note: The connecting lines between the items do not indicate a continuous measurement. They are supposed to increase legibility and spotting differences between qualitative profiles.

Yet, univariate ANOVAs revealed no significant differences in the participants' perceptions of the visualisations between groups.

Finally, many participants (N = 246) provided additional feedback in a text field. To these (N=193), codes could be assigned to password-related feedback. The remaining statements were too unspecific or concerned the study as such, e.g. "good study". Exploratorily categorizing the comments revealed some statements that might be applicable to all conditions. For example, many people were generally positive about the password feedback (N = 106), considered the password feedback to be helpful (N = 21), clear (N = 6), and informative (N = 17). On the other hand, a few people found the password feedback annoying (N = 10), "unsatisfactory" (N = 6), confusing (N = 5), unnecessary (N = 3), or contained too much text (N = 3).

In terms of condition-specific feedback, only in the simple nudge condition did participants mention mission requirements and suggestions (N = 5). Some examples of condition-specific comments are provided below:

- Compensation Nudge: "*I also used the days until I need to change the password rather than just the color rating. This was helpful because it gave me measurable feedback to how secure it was.*"
- Descriptive Norm Nudge: "*I've never seen a password screen that compared the strength of my password to other users. It really did motivate me to strengthen my password and get about 98%. I think that's a great idea.*"
- Fear Appeal Nudge: "*I loved the hacker stating how many years it would take to break in. lol.*"
- Information: "*I would find such password feedback helpful in real life, though it might be slightly annoying at times.* "

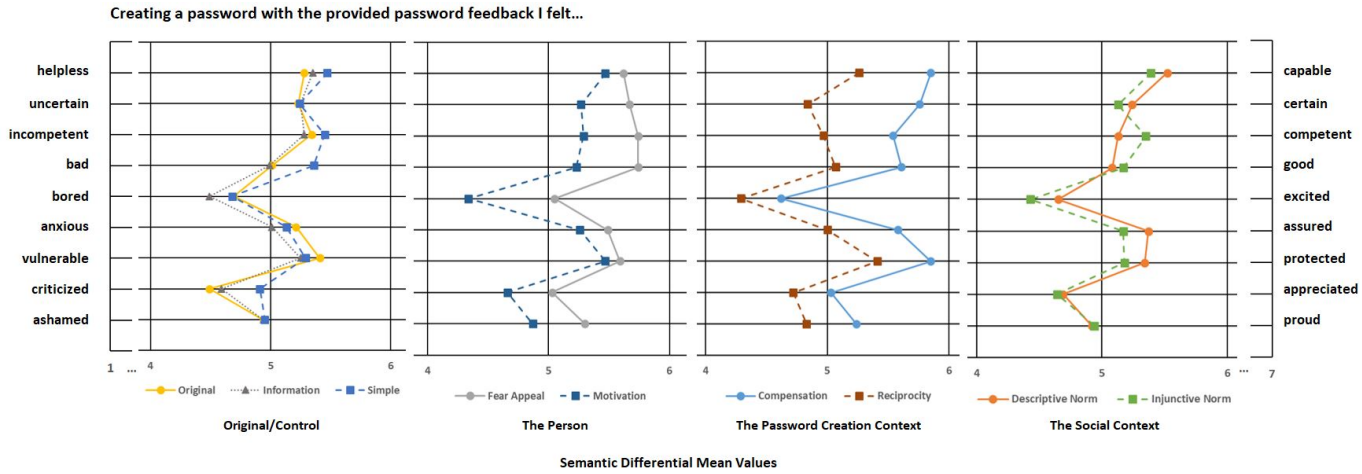

Figure 21. Profile diagrams of the user ratings' mean values in terms of how they felt creating a password with the provided feedback. Please note: The connecting lines between the items do not indicate a continuous measurement. They are supposed to increase legibility and spotting differences between qualitative profiles.

- Motivation Nudge: "*I thought it definitely encouraged me to create a stronger more complicated password than I normally would.*"

- Simple Nudge: "*Simply rating the password as good or bad (by the color bar) didn't teach me anything about how to improve my choice.* "

Control Variables

The sample's technological affinity, measured with the ATI scale [30] ranging from one to six, was $M = 4.01$ ($SD = .82$, $Md = 4$). The participants' security behaviour in terms of password creation measured with the SEBIS [26] centered around the middle of the five-point-scale with $M = 3.34$ ($SD = .83$, $Md = 3.25$).

On a five-point scale the participants' self-efficacy was reported as $M = 3.97$ ($SD = .72$, $Md = 4$).

DISCUSSION

The results partially confirm H_1, H_2, and H_3, suggesting that the combination of password information and a nudge within a password meter is more effective in increasing password strength, length, and entropy, as compared to either intervention on its own. Of the seven hybrid nudge password meters, including the original version by Ur *et al.* [109] and six variations thereof, four to five variations significantly outranked the simple nudge condition in terms of password strength, length, and entropy (see Table 3). The same is true for the comparison to the information condition.

Regarding RQ_1, considering differences in password creation between various hybrid nudge password meter variants, the results show that only the motivation nudge password meter exceeded the original hybrid nudge password meter in terms of improved strength, length, and entropy. In line with the finding that the motivation nudge encouraged at least equally strong passwords as the fear appeal nudge, previous research found inducing positive emotions to lead to stronger password creation, as compared to inducing negative emotions [41]. Another explanation of the difference might lie in the fear appeal targeting extrinsic motivation (driven by external rewards) [71], while the motivation nudge targets intrinsic motivation (driven by personal interest an engagement) that in some studies has been found to be more effective [10]. Furthermore, there are slight, but non-significant differences between the password meters targeting the person, the password creation context, or the social context.

The difference between the hybrid password meter variations, and in some cases between the hybrid and the control conditions 'simple nudge' and 'information,' were not significant. This, even though the tendency is visible in the descriptive values, might be due to the limited sample size within each group. Because a relatively large number of participants had to be excluded from the analysis after they admitted having reused an existing password or having used a password manager (even though instructed otherwise), the participant numbers in each group was reduced. It might well be that an existing but small effect was not detected. An indication that sample size might have played a role is given by the results of the study by Ur *et al.* [109] that found significant differences between the original hybrid password meter and the simple nudge condition with larger sample sizes.

Even so, in terms of password creation, as long as suitable information and a matching nudge are combined, it does not seem to make a huge difference what norm, bias or heuristic the nudge is based on or whether it targets the person, the password creation context or the social context. In line with that, some of the authors of previous research [38, 111, 109] came to the conclusion that the design of the feedback bar or other visualizations had little impact on password strength, but that "*an important factor seemed to be the combination of text and a visual indicator, rather than only having text or only having a visual bar*" [111] (p.14). Our research suggests that this conclusion can be extended to the type of bias, heuristic or

norm targeted by the visualization. One possible explanation for this finding might be that the dynamic, textual feedback and password guidance is already so "strong" or salient that in the *interaction* with other visual design elements the type of feedback nudge is of lower priority.

Because the sample sizes for comparing differences in terms of password memorability (RQ_2) were rather small, we can only conclude that the descriptive values provide no indication of lower password memorability of passwords created with a hybrid password meter, as compared to those only containing either information or a nudge. Memorability rates for the different hybrid password meter variations, even though the passwords were generally stronger, were mostly higher than those of the two control groups.

An unanticipated but interesting finding is that a large number of participants admitted to having recorded their password but nonetheless were not able to reproduce it correctly. It would be interesting to follow up this finding, the reasons for this failure and to determine whether the problem transfers to real-world passwords. It might well be a paper record of password is misplaced or misfiled, or that people had technical problems retrieving their password from browsers and/or password managers.

With respect to RQ_3, the password feedback overall was rated positively with rating means between 4 for the lowest-rated conditions and 6 on a scale from 1 to 7. Even so, looking at the descriptive values reveals other differences. For example, the password feedback by compensation nudge, fear appeal nudge and motivation nudge participants received the highest descriptive values on a number of items whereas the simple nudge and information condition received some of the lowest values.

The fact that, in line with the descriptive values, people perceived the motivation nudge passwords to be the strongest, and the simple nudge passwords as weakest can be seen as an indication that the password feedback did indeed support people in rating their password strength and in aligning user perceptions and actual password strength. This provides an indication that the user perceptions complement H_1, H_2, and H_3, in that hybrid password meters not only *support* stronger password creation but are also *viewed as supporting stronger password creation* and as being, e.g., more "helpful", "intuitive" and "fun" as compared to the single intervention password meters. The positive reception of the hybrid password meters suggests that the participants endorse this type of intervention which constitutes an important aspect in terms of the deployment of nudges [106].

Furthermore, the transparent design of the nudges offers recipients the chance to engage in the process and express concerns, as suggested by Sunstein *et al.* [106]. Suggestions for improving the password meter in this research include reducing the amount of text and addressing the sometimes perceived "unsatisfactory" nature of password meters, e.g. by adapting the password meter to the differing security levels required by different account types.

To summarize, the results provide support for the hypothesis that the combination of information and a feedback nudge is more effective in encouraging people to create a strong password than either intervention on its own. The user perceptions seem to provide further support for hybrid password meters that were generally rated more positively, and memorability rates were also higher rather than lower for passwords created when a hybrid password meter was present. Yet, it does not seem to make much of a difference which bias or heuristic the nudge is based on, as long as it is suitable and targeted to the context it is applied in. This study shows that the specific nudges designed to target the person, the password creation context, or the social context more or less affected password creation in an equally positive manner.

LIMITATIONS & FUTURE WORK

This section first details the limitations of the literature review before reflecting on the hybrid password meter study.

Literature Review

Even though a variety of measures were applied to include most relevant articles studying password meters, such as an additional forward and backward search, the literature review is probably not exhaustive and acts as a snapshot of the existing research. All included articles were carefully analysed. Nevertheless, it is possible that due to misunderstanding or a lack of information given the often concise format, some descriptions in the articles could have been misinterpreted and thus resulted in a different categorization than the authors of the papers would have selected themselves. The outcome of the literature review cannot be used to derive final conclusions, but serves as an overview of relevant, previous work and helped us to derive hypotheses and directions for future studies in this domain, as indeed we did ourselves.

Hybrid Password Meters Study

As a first step towards an empirical comparison of hybrid password meters that are systematically varied in one aspect, the type of feedback nudge, this research analysed password creation at one point in time. For future work it would be interesting to analyse whether differences arise in the long term. For example, it is possible that password creation under the compensation nudge linking password strength to password lifetime, where one could expect the effect to be especially notable in the long term, might be impacted differently, as compared to other nudges. It would be interesting to analyse whether the motivation nudge can maintain its strong influence on password creation or whether its novelty wears off over time. Furthermore, it would be relevant to analyse and confirm the external validity of the findings in real-life settings, including actual and relevant accounts even though previous research has shown that passwords created in role-playing tasks are indeed representative of real-world password choices [57, 28, 63].

In line with other research [94, 118, 74], our study shows that even though participants were instructed to create a new, not previously used password, passwords were reused in about 9% of cases. Password managers were used in about 10% of cases, which slightly exceeds previous findings, e.g. [83, 61]. While reuse might be attributed to the artificial study context (even

though also found in real life), the use of password managers might be more prevalent amongst MTurk workers who might be younger and/or more tech-savvy. Future research could analyse the acceptance of a combination of the password creation tool tested in this research and password managers. For example, it could be suggested that people store their passwords in a password manager linked to the tool to eliminate memorability concerns. Currently, password managers are not wide-spread [61, 75, 83] and, among other things, suffer from trust issues [7, 75]. It would be interesting to see whether a combination of tools would increase acceptance.

Password memorability rates equalled about one third of the people who returned for the follow-up study and was thus lower than the rates found by Ur *et al.* [109]. This can be explained by the much longer time delay between the first the follow-up studies, as compared to Ur *et al.'s* 48 hours. While the longer duration might be more helpful in analysing long-term memorability, an additional data point after 48 hours might have helped to gain a clearer picture of password memorability across the different conditions and encouraged higher return rates.

In this study, different password strength scores, memorability, and the users' subjective assessment were evaluated. It would be beneficial to analyse how long it took the participants to create strong passwords with the different hybrid password meters and whether or how this affected their subjective evaluation. Future work should thus consider timing data as a relevant measure.

Aiming to recruit a large enough number of participants to be able to detect potential differences between the numerous conditions, we opted for an online study. For future work, however, it would be interesting to analyse a smaller number of conditions in more detail in a laboratory study. It would then be feasible to explore the users' perceptions in a follow-up interview or to use eye-tracking to analyse which elements of the hybrid password meters participants focus on more than others. These objective measures could well support and enhance the subjective ratings we collected with regards to the individual design elements included in the hybrid password meters.

CONCLUSION

Commencing with a literature review into password meters, we found that studies more often found password meters to be effective than not, across a range of contexts and samples, and that their presence did not degrade password memorability. This suggests that password meters are indeed a promising approach and reveals the potential of softly guiding users towards secure passwords as an alternative to enforcement of strict rules. Furthermore, besides strength feedback, effective password meters often included: (a) a feedback nudge encouraging users to increase password strength, and (b) additional information on how to create strong passwords. Yet, given the large number of differences in the study methodologies and password meter designs, both findings warranted further analysis: first, in terms of the impact of feedback nudge design, and second, in terms of the combined versus single impact of interventions on password creation.

We set out to explore whether the design of the feedback nudges included in password meters influenced the impact of the password meter in different ways. We tested multiple variants of hybrid password meters, combining password information and a variety of feedback nudges, based on a successfully evaluated password meter [109]. These were developed and compared in a between-subjects user study in terms of password strength, password memorability, and participants' perceptions. Two password meter variants using either a feedback nudge (a colored bar) or password information served as a comparison for the single versus combined impact of the interventions. The findings confirmed the insight gained from the literature review, i.e., *that the combination of information and a feedback nudge was more successful in encouraging strong passwords than either intervention on its own*. Furthermore, even though produced passwords were generally stronger, password memorability was not negatively impacted. Participants favored the hybrid password meters finding them to be more helpful, clear, and fun to use.

The different hybrid password meters led to slight but insignificant differences, except for one variation. With larger sample sizes small effects might emerge, but the differences do not seem to be huge. This leads to the conclusion that it might be more important to combine suitable password information and feedback nudges targeted at the given context, as compared to the choice of bias, heuristic, or norm the nudge is based on. Future work should analyse potential differences and habituation effects in the long-term, and analyse the external validity of the findings in the field, i.e. in real-life settings where the password is required to access something the participant cares about.

Overall, the findings lead to the following six recommendations:

1. **Support Password Creation:** Service-providers and developers should consider supporting the password creation process by implementing a password meter that provides feedback on the users' password strength.

2. **Encourage Users with Nudges:** When implementing a password meter, include a nudge encourage users to increase password strength, together with information on how to achieve this.

3. **Consider the Impact of Nudge Designs:** This research suggests that the type of nudge plays a secondary role as long as it is suitable for the targeted user group and context. Even so, any nudge should be carefully designed and evaluated in terms of its effectiveness and potential unanticipated side-effects to ensure ethical deployment.

4. **Provide Transparent Nudges:** The nudge should be transparent to the user. For example, all nudge visualizations and instructions used in this study were designed to be visible and understandable for the users. Instead, the exclusive use of a framing nudge in the instruction, without further reference or explanation, might go undetected.

5. **Consider the Combination with Password Managers:** Future research might consider combining a password meter

with a built-in password manager that stores the user-created passwords to eliminate memorability considerations. However, the current lack of trust in password managers, potential downsides such as a password manager being a single point of failure, and the responsibility resulting from suggesting a certain tool will have to be acknowledged.

DECLARATION OF INTEREST STATEMENT

There are no relevant financial or non-financial competing interests to report.

Word Count: Ca. 14.000 words not including references or the appendix.

REFERENCES

[1] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, and others. 2017. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys (CSUR)* 50, 3 (2017), 1–41.

[2] Anne Adams and Martina Angela Sasse. 1999. Users are not the enemy. *Commun. ACM* 42, 12 (1999), 41–46.

[3] Anne Adams, Martina Angela Sasse, and Peter Lunt. 1997. Making passwords secure and usable. In *People and Computers XII*, Harold Thimbleby, Brid O'Conaill, and Peter J. Thomas (Eds.). Springer, Bristol, England, UK, 1–19.

[4] Nouf Aljaffan, Haiyue Yuan, and Shujun Li. 2017. PSV (Password Security Visualizer): From Password Checking to User Education. In *Proceedings of the International Conference on Human Aspects of Information Security, Privacy, and Trust*. Springer, Cham, Switzerland, 191–211.

[5] Hunt Allcott and Sendhil Mullainathan. 2010. Behavior and energy policy. *Science* 327, 5970 (2010), 1204–1205.

[6] Nalin Asanka Gamagedara Arachchilage and Steve Love. 2013. A game design framework for avoiding phishing attacks. *Computers in Human Behavior* 29, 3 (2013), 706–714.

[7] Salvatore Aurigemma, Thomas Mattson, and Lori Leonard. 2017. So much promise, so little use: What is stopping home end-users from using password manager applications?. In *Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS 2017)*. Association for Information Systems, Atlanta, USA, 4061–4070.

[8] Ingolf Becker, Simon Parkin, and M Angela Sasse. 2018. The rewards and costs of stronger passwords in a university: linking password lifetime to strength. In *Proceedings of the 27th USENIX Security Symposium*. USENIX, Berkely, CA, USA, 239–253.

[9] Constanze Beierlein, Anastassiya Kovaleva, Christoph J Kemper, and Beatrice Rammstedt. 2012. Ein Messinstrument zur Erfassung subjektiver Kompetenzerwartungen: Allgemeine Selbstwirksamkeit Kurzskala (ASKU). *GESIS-Working Papers* 2012/17 (2012), 1–24.

[10] Roland Benabou and Jean Tirole. 2003. Intrinsic and extrinsic motivation. *The Review of Economic Studies* 70, 3 (2003), 489–520.

[11] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300.

[12] George Y Bizer, Rachel A Magin, and Madeline R Levine. 2014. The Social-Norm Espousal Scale. *Personality and Individual Differences* 58 (2014), 106–111.

[13] Joseph Bonneau, Cormac Herley, Paul C Van Oorschot, and Frank Stajano. 2012. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*. IEEE, New York, NY, USA, 553–567.

[14] William E Burr, Donna F Dodson, Elaine M Newton, Ray A Perlner, W Timothy Polk, Sarbari Gupta, and Emad A Nabbus. 2011. National Institute of Standards & Technology Sp 800-63-1. Electronic Authentication Guideline. (2011).

[15] Ryan Calo. 2014. Code, Nudge or Notice? *Iowa Law Review* 99 (2014), 773–802.

[16] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 503.1–503.15.

[17] Sonia Chiasson and Paul C Van Oorschot. 2015. Quantifying the security advantage of password expiration policies. *Designs, Codes and Cryptography* 77, 2-3 (2015), 401–408.

[18] Eun Kyoung Choe, Jaeyeon Jung, Bongshin Lee, and Kristie Fisher. 2013. Nudging people away from privacy-invasive mobile apps through visual framing. In *Proceedings of the IFIP Conference on Human-Computer Interaction*. Springer, Berlin/Heidelberg, Germany, 74–91.

[19] Adrienne Chung and Rajiv N Rimal. 2016. Social norms: A review. *Review of Communication Research* 4 (2016), 1–28.

[20] Robert B Cialdini, Raymond R Reno, and Carl A Kallgren. 1990. A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology* 58, 6 (1990), 1015–1026.

[21] Mark Ciampa. 2013. A comparison of password feedback mechanisms and their impact on password entropy. *Information Management & Computer Security* 21, 5 (2013), 344–359.

[22] Xavier de Carné de Carnavalet and Mohammad Mannan. 2014. From very weak to very strong: Analyzing password-strength meters. In *Network and Distributed System Security Symposium (NDSS 2014)*. Internet Society, Reston, VA, USA, 1–16.

[23] Matteo Dell'Amico and Maurizio Filippone. 2015. Monte Carlo strength evaluation: Fast and reliable password checking. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, New York, NY, USA, 158–169.

[24] Marc Dupuis and Faisal Khan. 2018. Effects of peer feedback on password strength. In *Proceedings of the 2018 APWG Symposium on Electronic Crime Research (eCrime)*. IEEE, New York, NY, USA, 1–9.

[25] David Eargle, John Godfrey, Hsin Miao, Scott Stevenson, Rich Shay, Blase Ur, and Lorrie Cranor. 2015. Poster: You can do better–motivational statements in password-meter feedback. In *Proceedings of the 11th Symposium On Usable Privacy and Security (SOUPS)*. Usenix, Berkeley, CA, USA, 1–2.

[26] Serge Egelman and Eyal Peer. 2015. Scaling the security wall: Developing a security behavior intentions scale (sebis). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2873–2882.

[27] Serge Egelman, Andreas Sotirakopoulos, Ildar Muslukhov, Konstantin Beznosov, and Cormac Herley. 2013. Does my password go up to eleven? The impact of password meters on password selection. In *Proceedings of the 2013 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2379–2388.

[28] Sascha Fahl, Marian Harbach, Yasemin Acar, and Matthew Smith. 2013. On the ecological validity of a password study. In *Proceedings of the 9th Symposium on Usable Privacy and Security*. ACM, New York, NY, USA, 1–13.

[29] Dinei Florencio and Cormac Herley. 2007. A large-scale study of web password habits. In *Proceedings of the 16th International Conference on World Wide Web*. ACM, New York, NY, USA, 657–666.

[30] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467.

[31] Steven Furnell. 2011. Assessing password guidance and enforcement on leading websites. *Computer Fraud & Security* 2011, 12 (2011), 10–18.

[32] Steven Furnell, Faisal Alotaibi, and Rawan Esmael. 2019. Aligning security practice with policy: Guiding and nudging towards better behavior. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*. Scholar Space, Honolulu, Hawaii, USA, 5618–5627.

[33] Steven Furnell and Nina Bär. 2013. Essential lessons still not learned? Examining the password practices of end-users and service providers. In *Proceedings of the International Conference on Human Aspects of Information Security, Privacy, and Trust*. Springer, Cham, Switzerland, 217–225.

[34] Steven Furnell, Warut Khern-am nuai, Rawan Esmael, Weining Yang, and Ninghui Li. 2018. Enhancing security behaviour by supporting the user. *Computers & Security* 75 (2018), 1–9.

[35] Nina Gerber, Benjamin Reinheimer, and Melanie Volkamer. 2019. Investigating People's Privacy Risk Perception. *Proceedings on Privacy Enhancing Technologies* 2019, 3 (2019), 267–288.

[36] Richard J Gerrig and PG Zimbardo. 2014. *Psychologie (20. Aufl.)*. Pearson, Hallbergmoos, Germany.

[37] Maximilian Golla and Markus Dürmuth. 2018. On the accuracy of password strength meters. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. ACM, New York, NY, USA, 1567–1582.

[38] Maximilian Golla, Björn Hahn, Karsten Meyer zu Selhausen, Henry Hosseini, and Markus Dürmuth. 2018. Bars, Badges, and High Scores: On the Impact of Password Strength Visualizations. In *Proceedings of Who Are You?! Adventures in Authentication (WAY)*. Usenix, Berkeley, CA, USA, 1–7.

[39] Paul A. Grassi, James L. Fenton, and Michael E. Garcia. 2017. *Digital Identity Guidelines [including updates as of 12-01-2017]*. Technical Report. NIST Special Publication 800-63-3, Gaithersburg, MD, USA.

[40] Beate Grawemeyer and Hilary Johnson. 2011. Using and managing multiple passwords: A week to a view. *Interacting With Computers* 23, 3 (2011), 256–267.

[41] Iwan Gulenko. 2014. Improving passwords: influence of emotions on security behaviour. *Information Management & Computer Security* 22, 2 (2014), 167–178.

[42] Hana Habib, Jessica Colnago, William Melicher, Blase Ur, Sean Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Cranor. 2017. Password creation in the presence of blacklists. In *Proceedings of the Workshop on Usable Security and Privacy (USEC)*. Internet Society, Reston, VA, USA, 1–11.

[43] Hana Habib, Pardis Emami Naeini, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2018. User behaviors and attitudes under password expiration policies. In *Proceedings of the 14th Symposium on Usable Privacy and Security (SOUPS)*. Usenix, Berkeley, CA, USA, 13–30.

[44] Pelle Guldborg Hansen. 2016. The definition of nudge and libertarian paternalism: Does the hand fit the glove? *European Journal of Risk Regulation* 7, 1 (2016), 155–174.

[45] Pelle Guldborg Hansen and Andreas Maaløe Jespersen. 2013. Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation* 4, 1 (2013), 3–28.

[46] Marc Hassenzahl, Michael Burmester, and Franz Koller. 2003. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Proceedings of the Mensch & Computer 2003*. Vieweg+Teubner Verlag, Wiesbaden, Germany, 187–196.

[47] Daniel M Hausman and Brynn Welch. 2010. Debate: To nudge or not to nudge. *Journal of Political Philosophy* 18, 1 (2010), 123–136.

[48] Susanna Heidt and Adam J Aviv. 2016. Refining Graphical Password Strength Meters for Android Phones. In *Poster presented at the 12th Symposium on Usable Security and Privacy (SOUPS)*. Usenix, Berkely, CA, USA, 1–5.

[49] Cormac Herley and Paul Van Oorschot. 2011. A research agenda acknowledging the persistence of passwords. *IEEE Security & Privacy* 10, 1 (2011), 28–36.

[50] Philip G. Inglesant and M. Angela Sasse. 2010. The True Cost of Unusable Password Policies: Password Use in the Wild. In *Proceedings of the 2010 CHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 383–392. DOI: http://dx.doi.org/10.1145/1753326.1753384

[51] Christina Katsini, Marios Belk, Christos Fidas, Nikolaos Avouris, and George Samaras. 2016. Security and usability in knowledge-based user authentication: A review. In *Proceedings of the 20th Pan-Hellenic Conference on Informatics*. ACM, New York, NY, USA, 1–6.

[52] Patrick Gage Kelley, Saranga Komanduri, Michelle L Mazurek, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Julio Lopez. 2012. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *Proceedings of the IEEE Symposium on Security and Privacy*. IEEE, New York, NY, USA, 523–537.

[53] Warut Khern-am nuai, Matthew J Hashim, Alain Pinsonneault, Weining Yang, and Ninghui Li. 2019. Enhancing Operational Security by Redesigning Password Strength Meters: Evidence from Randomized Experiments. (2019). Available at SSRN 2800499.

[54] Warut Khern-am nuai, Weining Yang, and Ninghui Li. 2017. Using Context-Based Password Strength Meter to Nudge Users' Password Generating Behavior: A Randomized Experiment. In *Proceedings of the 50th Hawaii International Conference on System Sciences*. AIS Electronic Library (AISeL), Big Island, Hawaii, 587–596.

[55] Tiffany Hyun-Jin Kim, H Colleen Stuart, Hsu-Chun Hsiao, Yue-Hsun Lin, Leon Zhang, Laura Dabbish, and Sara Kiesler. 2014. YourPassword: applying feedback loops to improve security behavior of managing multiple passwords. In *Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security*. ACM, New York, NY, USA, 513–518.

[56] Saranga Komanduri, Richard Shay, Lorrie Faith Cranor, Cormac Herley, and Stuart Schechter. 2014. Telepathwords: Preventing weak passwords by reading users' minds. In *Proceedings of the 23rd USENIX Security Symposium*. Usenix, Berkeley, CA, USA, 591–606.

[57] Saranga Komanduri, Richard Shay, Patrick Gage Kelley, Michelle L Mazurek, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Serge Egelman. 2011. Of passwords and people: measuring the effect of password-composition policies. In *Proceedings of the 2011 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2595–2604.

[58] Floor M Kroese, David R Marchiori, and Denise TD de Ridder. 2015. Nudging healthy food choices: a field experiment at the train station. *Journal of Public Health* 38, 2 (2015), e133–e137.

[59] Dhananjay Kulkarni. 2010. A novel web-based approach for balancing usability and security requirements of text passwords. *International Journal of Network Security & its Applications (0975-2307)* 2, 3 (2010), 1–16.

[60] Huigang Liang and Yajiong Xue. 2010. Understanding security behaviors in personal computer usage: A threat avoidance perspective. *Journal of the Association for Information Systems* 11, 7 (2010), 394–413.

[61] Sanam Ghorbani Lyastani, Michael Schilling, Sascha Fahl, Michael Backes, and Sven Bugiel. 2018. Better managed than memorized? Studying the Impact of Managers on Password Strength and Reuse. In *Proceedings of the 27th USENIX Security Symposium*. Usenix, Berkeley, CA, USA, 203–220.

[62] Jerry Ma, Weining Yang, Min Luo, and Ninghui Li. 2014. A study of probabilistic password models. In *Proceedings of the 35th IEEE Symposium on Security and Privacy*. IEEE, New York, NY, USA, 689–704.

[63] Michelle L Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Patrick Gage Kelley, Richard Shay, and Blase Ur. 2013. Measuring password guessability for an entire university. In *Proceedings of the ACM SIGSAC Conference on Computer & Communications Security*. ACM, New York, NY, USA, 173–186.

[64] Adam W Meade and S Bartholomew Craig. 2012. Identifying careless responses in survey data. *Psychological Methods* 17, 3 (2012), 437.

[65] William Melicher, Darya Kurilova, Sean M Segreti, Pranshu Kalvani, Richard Shay, Blase Ur, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Michelle L Mazurek. 2016. Usability and security of text passwords on mobile devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 527–539.

[66] William Melicher, Blase Ur, Sean M Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2017. Better passwords through science (and neural networks). *Usenix ;login:* 42, 4 (2017), 26–30.

[67] William Melicher, Blase Ur, Sean M Segreti, Saranga Komanduri, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2016. Fast, lean, and accurate: Modeling password guessability using neural networks. In *Proceedings of the 25th USENIX Security Symposium (USENIX Security)*. Usenix, Berkeley, CA, USA, 175–191.

[68] K Mohamed. 2014. Password-Meter-Tutorial. (2014). GitHub.

[69] European Federation of Psychologists' Association. 2005. Meta-Code of Ethics. (2005). `https://www.bdp-verband.de/binaries/content/assets/beruf/efpa_metacode_en.pdf`.

[70] Takahiro Ohyama and Akira Kanaoka. 2015. Poster: Password Strength Meters using Social Influence. In *Proceedings of the 11th Symposium On Usable Privacy and Security (SOUPS)*, Vol. 2015. Usenix, Berkeley, CA, USA, 1–2.

[71] Jacques Ophoff and Frauke Dietz. 2019. Using Gamification to Improve Information Security Behavior: A Password Strength Experiment. In *Proceedings of the IFIP World Conference on Information Security Education*. Springer, Cham, Switzerland, 157–169.

[72] Ronald Paans and IS Herschberg. 1987. Computer security: the long road ahead. *Computers & Security* 6, 5 (1987), 403–416.

[73] Bijeeta Pal, Tal Daniel, Rahul Chatterjee, and Thomas Ristenpart. 2019. Beyond credential stuffing: Password similarity models using neural networks. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*. IEEE, New York, NY, USA, 417–434.

[74] Sarah Pearman, Jeremy Thomas, Pardis Emami Naeini, Hana Habib, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, and Alain Forget. 2017. Let's go in for a closer look: Observing passwords in their natural habitat. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, New York, NY, USA, 295–310.

[75] Sarah Pearman, Shikun Aerin Zhang, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2019. Why people (don't) use password managers effectively. In *Fifteenth Symposium on Usable Privacy and Security ({SOUPS} 2019)*. USENIX Association, Berkeley, CA, USA, 319–338.

[76] Eyal Peer, Serge Egelman, Marian Harbach, Nathan Malkin, Arunesh Mathur, and Alisa Frik. 2020. Nudge me right: Personalizing online security nudges to people's decision-making styles. *Computers in Human Behavior* 109 (2020), 106347.1–106347.9.

[77] Imran Rasul and David Hollywood. 2012. Behavior change and energy use: is a 'nudge' enough? *Carbon Management* 3, 4 (2012), 349–351.

[78] Karen Renaud and Marc Dupuis. 2019. Cyber Security Fear Appeals: Unexpectedly Complicated. In *Proceedings of the 2019 New Security Paradigms Workshop (NSPW)*. ACM, New York, NY, USA, 1–15.

[79] Karen Renaud and Verena Zimmerman. 2017. Enriched nudges lead to stronger password replacements... but implement mindfully. In *Proceedings of the Information Security for South Africa (ISSA) conference*. IEEE, New York, NY, USA, 1–9.

[80] Karen Renaud, Verena Zimmerman, Joseph Maguire, and Steve Draper. 2017. Lessons learned from evaluating eight password nudges in the wild. In *Proceedings of the The LASER Workshop: Learning from Authoritative Security Experiment Results*. Usenix, Berkeley, CA, USA, 25–37.

[81] Karen Renaud and Verena Zimmermann. 2018a. Ethical guidelines for nudging in information security & privacy. *International Journal of Human-Computer Studies* 120 (2018), 22–35.

[82] Karen Renaud and Verena Zimmermann. 2018b. Guidelines for ethical nudging in password authentication. *SAIEE Africa Research Journal* 109, 2 (2018), 102–118.

[83] Karen Renaud and Verena Zimmermann. 2019a. Encouraging password manager use. *Network Security* 2019, 6 (2019), 20–20.

[84] Karen Renaud and Verena Zimmermann. 2019b. Nudging folks towards stronger password choices: providing certainty is the key. *Behavioural Public Policy* 3, 2 (2019), 228–258.

[85] Przemysław Rodwald. 2019. Using gamification and fear appeal instead of password strength meters to increase password entropy. *Scientific Journal of Polish Naval Academy* 217, 2 (2019), 17–33.

[86] Scott Ruoti, Jeff Andersen, and Kent Seamons. 2016. Strengthening password-based authentication. In *Proceedings of the 12th Symposium on Usable Privacy and Security (SOUPS)*. Usenix, Berkely, CA, USA, 1–2.

[87] Karen Scarfone and Murugiah Souppaya. 2009. *Guide to enterprise password management*. Technical Report. National Institute of Standards and Technology.

[88] P Wesley Schultz, Jessica M Nolan, Robert B Cialdini, Noah J Goldstein, and Vladas Griskevicius. 2007. The constructive, destructive, and reconstructive power of social norms. *Psychological Science* 18, 5 (2007), 429–434.

[89] Sean M Segreti, William Melicher, Saranga Komanduri, Darya Melicher, Richard Shay, Blase Ur, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Michelle L Mazurek. 2017. Diversify to survive: making passwords stronger with adaptive policies. In *Proceedings of the 13th Symposium on Usable Privacy and Security (SOUPS)*. Usenix, Berkeley, CA, USA, 1–12.

[90] Tobias Seitz and Heinrich Hussmann. 2017. PASDJO: quantifying password strength perceptions with an online game. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction (OzCHI)*. ACM, New York, NY, USA, 117–125.

[91] Tobias Seitz, Emanuel von Zezschwitz, Stefanie Meitner, and Heinrich Hussmann. 2016. Influencing Self-Selected Passwords Through Suggestions and the Decoy Effect. In *Proceedings of the 1st European Workshop on Usable Security*. Internet Society, Reston, VA, USA, 1–7.

[92] Claude E Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27, 3 (1948), 379–423.

[93] Richard Shay, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Alain Forget, Saranga Komanduri, Michelle L Mazurek, William Melicher, Sean M Segreti, and Blase Ur. 2015. A spoonful of sugar? The impact of guidance and feedback on password-creation behavior. In *Proceedings of the 33rd CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2903–2912.

[94] Richard Shay, Saranga Komanduri, Patrick Gage Kelley, Pedro Giovanni Leon, Michelle L Mazurek, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2010. Encountering stronger password requirements: user attitudes and behaviors. In *Proceedings of the 6th Symposium on Usable Privacy and Security*. ACM, New York, NY, USA, 1–20.

[95] Lynsay A Shepherd and Jacqueline Archibald. 2017. Security awareness and affective feedback: categorical behaviour vs. reported behaviour. In *Proceedings of the International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)*. IEEE, New York, NY, USA, 1–6.

[96] Lynsay A Shepherd, Jacqueline Archibald, and Robert Ian Ferguson. 2017. Assessing the impact of affective feedback on end-user security awareness. In *Proceedings of the International Conference on Human Aspects of Information Security, Privacy, and Trust*. Springer, Cham, Switzerland, 143–159.

[97] Kamran Siddique, Zahid Akhtar, and Yangwoo Kim. 2017. Biometrics vs passwords: a modern version of the tortoise and the hare. *Computer Fraud & Security* 2017, 1 (2017), 13–17.

[98] Youngbae Song, Geumhwan Cho, Seongyeol Oh, Hyoungshick Kim, and Jun Ho Huh. 2015. On the effectiveness of pattern lock strength meters: Measuring the strength of real world pattern locks. In *Proceedings of the 33rd CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2343–2352.

[99] Andreas Sotirakopoulos, Ildar Muslukov, Konstantin Beznosov, Cormac Herley, and Serge Egelman. 2011. Poster: Motivating users to choose better passwords through peer pressure. In *Proceedings of the 7th Symposium On Usable Privacy and Security (SOUPS)*, Vol. 2011. ACM, New York, NY, USA, 1–2.

[100] Michael Stainbrook and Nicholas Caporusso. 2019. Comparative Evaluation of Security and Convenience Trade-Offs in Password Generation Aiding Systems. In *Proceedings of the International Conference on Applied Human Factors and Ergonomics*. Springer, Cham, Switzerland, 87–96.

[101] Statista. 2018. Cybersecurity & Cloud 2018. (August 2018). `https://de.statista.com/statistik/studie/id/58204/dokument/cybersecurity-und-cloud/` (accessed 09 June 2020).

[102] Elizabeth Stobert and Robert Biddle. 2014. The password life cycle: user behaviour in managing passwords. In *Proceedings of the Tenth Symposium On Usable Privacy and Security (SOUPS)*. Usenix, Berkeley, CA, USA, 243–255.

[103] Taku Sugai, Toshihiro Ohigashi, Yoshio Kakizaki, and Akira Kanaoka. 2019. Password Strength Measurement without Password Disclosure. In *Proceedings of the 14th Asia Joint Conference on Information Security (AsiaJCIS)*. IEEE, New York, NY, USA, 157–164.

[104] Chen Sun, Yang Wang, and Jun Zheng. 2014. Dissecting pattern unlock: The effect of pattern strength meter on pattern selection. *Journal of Information Security and Applications* 19, 4-5 (2014), 308–320.

[105] Cass R Sunstein. 1996. Social norms and social roles. *Columbia Law Review* 96, 4 (1996), 903–968.

[106] Cass R Sunstein, Lucia A Reisch, and Micha Kaiser. 2019. Trusting nudges? Lessons from an international survey. *Journal of European Public Policy* 26, 10 (2019), 1417–1443.

[107] Richard H Thaler and Cass R. Sunstein. 2008. *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press, New Haven, CT, US.

[108] Christian Tiefenau, Maximilian Häring, Mohamed Khamis, and Emanuel von Zezschwitz. 2019. "Please enter your PIN"–On the Risk of Bypass Attacks on Biometric Authentication on Mobile Devices. (2019). arXiv preprint arXiv:1911.07692.

[109] Blase Ur, Felicia Alfieri, Maung Aung, Lujo Bauer, Nicolas Christin, Jessica Colnago, Lorrie Faith Cranor, Henry Dixon, Pardis Emami Naeini, Hana Habib, and others. 2017. Design and evaluation of a data-driven password meter. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 3775–3786.

[110] Blase Ur, Jonathan Bees, Sean M Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2016. Do users' perceptions of password security match reality?. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 3748–3760.

[111] Blase Ur, Patrick Gage Kelley, Saranga Komanduri, Joel Lee, Michael Maass, Michelle L Mazurek, Timothy Passaro, Richard Shay, Timothy Vidas, and Lujo Bauer. 2012a. How does your password measure up? the effect of strength meters on password creation. In *Proceedings of the 21st USENIX Security Symposium*. Usenix, Berkeley, CA, USA, 65–80.

[112] Blase Ur, Patrick Gage Kelley, Saranga Komanduri, Joel Lee, Michael Maass, Michelle L. Mazurek, Timothy Passaro, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, and Julio Lopez. 2012b. Helping users create better passwords. *USENIX ;login:Magazine* 37 (2012), 51–57. Issue 6.

[113] Blase Ur, Fumiko Noma, Jonathan Bees, Sean M Segreti, Richard Shay, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2015a. "I Added'!'at the End to Make It Secure": Observing Password Creation in the Lab. In *Proceedings of the 11th Symposium On Usable Privacy and Security (SOUPS)*. Usenix, Berkeley, CA, USA, 123–140.

[114] Blase Ur, Sean M Segreti, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Saranga Komanduri, Darya Kurilova, Michelle L Mazurek, William Melicher, and Richard Shay. 2015b. Measuring real-world accuracies and biases in modeling password guessability. In *Proceedings of the 24th USENIX Security Symposium*. Usenix, Berkeley, CA, USA, 463–481.

[115] Steven Van Acker, Daniel Hausknecht, Wouter Joosen, and Andrei Sabelfeld. 2015. Password meters and generators on the web: From large-scale empirical study to getting it right. In *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*. ACM, New York, NY, USA, 253–262.

[116] Anthony Vance, David Eargle, Kirk Ouimet, and Detmar Straub. 2013. Enhancing password security through interactive fear appeals: A web-based field experiment. In *Proceedings of the 46th Hawaii International Conference on System Sciences*. IEEE, New York, NY, USA, 2988–2997.

[117] Emanuel Von Zezschwitz, Alexander De Luca, and Heinrich Hussmann. 2013. Survival of the shortest: A retrospective analysis of influencing factors on password composition. In *Proceedings of the IFIP Conference on Human-Computer Interaction*. Springer, Berlin/Heidelberg, Germany, 460–467.

[118] Rick Wash, Emilee Rader, Ruthie Berman, and Zac Wellmer. 2016. Understanding password choices: How frequently entered passwords are re-used across websites. In *Proceedings of the 12th Symposium on Usable Privacy and Security (SOUPS)*. Usenix, Berkely, CA, USA, 175–188.

[119] David Watson and Lee Anna Clark. 1999. The PANAS-X: Manual for the positive and negative affect schedule-expanded form. *Iowa Research Online - The University of Iowa's Institutional Repository* (1999), 1–24. https://doi.org/10.17077/48vt-m4t2.

[120] Matt Weir, Sudhir Aggarwal, Michael Collins, and Henry Stern. 2010. Testing metrics for password creation policies by attacking large sets of revealed passwords. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*. ACM, New York, NY, USA, 162–175.

[121] Matt Weir, Sudhir Aggarwal, Breno De Medeiros, and Bill Glodek. 2009. Password cracking using probabilistic context-free grammars. In *Proceedings of the 30th IEEE Symposium on Security and Privacy*. IEEE, New York, NY, USA, 391–405.

[122] Daniel Lowe Wheeler. 2016. zxcvbn: Low-budget password strength estimation. In *Proceedings of the 25th USENIX Security Symposium*. Usenix, Berkeley, CA, USA, 157–173.

[123] Alma Whitten and J Doug Tygar. 1999. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0.. In *Proceedings of the USENIX Security Symposium*, Vol. 348. USENIX, Berkeley, CA, USA, 169–184.

[124] Kim Witte. 1992. Putting the fear back into fear appeals: The extended parallel process model. *Communication Monographs* 59, 4 (1992), 329–349. DOI:http://dx.doi.org/10.1080/03637759209376276

[125] Simon S Woo and Jelena Mirkovic. 2018. GuidedPass: Helping Users to Create Strong and Memorable Passwords. In *Proceedings of the International Symposium on Research in Attacks, Intrusions, and Defenses*. Springer, Cham, Switzerland, 250–270.

[126] Ming Xu and Weili Han. 2019. An Explainable Password Strength Meter Addon via Textual Pattern Recognition. *Security and Communication Networks* 2019 (2019), 1–10.

[127] Yinqian Zhang, Fabian Monrose, and Michael K Reiter. 2010. The security of modern password expiration: An algorithmic framework and empirical analysis. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*. ACM, New York, NY, USA, 176–186.

[128] Moshe Zviran and William J Haga. 1999. Password security: an empirical study. *Journal of Management Information Systems* 15, 4 (1999), 161–185.

APPENDIX

Appendix A: List of journals and conferences included in the literature review

Appendix B: Outcome of the literature review

The following table summarizes the results of the literature review including 42 articles. Articles referring to the same study were combined leading to 33 table entries.

Number	Conference/Journal	Number of articles identified	Number of articles included
	Google Scholar Rating		
1	ACM CHI Conference on Human Factors in Computing Systems (CHI)	23	4
2	International Journal of Human-Computer Studies (IJHCS)	1	0
3	IEEE Transactions on Human-Machine Systems	0	0
4	International Journal of Human-Computer Interaction	0	0
5	ACM Transactions on Computer-Human Interaction (TOCHI)	0	0
6	Human-centric Computing and Information Sciences (HCIS)	0	0
7	Annual Symposium on Computer-Human Interaction in Play (CHI PLAY)	0	0
8	Nordic Conference on Human-Computer Interaction (NordiCHI)	0	0
9	Australian Computer-Human Interaction Conference (OzCHI)	2	0
10	IFIP Conference on Human-Computer Interaction (INTERACT)	3	0
	Additions		
11	Annual Computer Security Applications Conference (ACSAC)	9	0
12	Computers in Human Behavior	4	0
13	Conference on Computer and Communications Security (CCS)	12	0
14	European Workshop on Usable Security (EuroUSEC)	3	1
15	Human-Computer Interaction (HCI)	0	0
16	Information and Computer Security (ICS)	6	2
17	Network and Distributed System Security Symposium (NDSS)	4	0
18	Passwords*	1	0
19	Symposium on Usable Privacy and Security (SOUPS)	14	2
20	Symposium on Security and Privacy (S&P)	6	0
21	USENIX Security	9	3
22	Workshop on Usable Security (USEC)	4	1
		100	**13**
	Forward and Backward Analysis		
	Resulting articles (N = 13)	107	29
	Overall	**207**	**42**

Table 4. Journals and Conferences included in the Literature Review along with the number of identified articles including the search terms and the number of included articles also meeting the inclusion criteria. *Only the 2014 and 2015 proceedings were available.

Appendix C: List of questionnaires and questionnaire items used in the study.

User perceptions of the password meter

The user perceptions of the password meter were measured on three sub scales that were inspired by the questions asked by Ur *et al.* [109] and the AttrakDiff [46] using a semantic differential on a 7-point scale:

Evaluation of the password creation process

Creating a password with the provided password feedback was...

- annoying - fun
- difficult - easy
- unintuitive - intuitive
- unpredictable - predictable
- conventional - novel
- confusing - clear
- discouraging - motivating
- trivial - challenging
- obstructive - helpful
- uninformative - informative

Evaluation of the created password

The password I created was...

- weak - strong
- simple - complex
- hard to remember - easy to remember
- short - long
- insecure - secure
- worse compared to other people's passwords - better compared to other people's passwords
- worse than passwords I usually use - better than passwords I usually use
- unimaginative - creative
- predictable - random

Author	No of PWMs	Type of PWM	Design Elements	Setting	Study Design	Participants	Security Measures	Comparison	PW strength	Other outcomes
Aljaffan, Yuan & Li (2017) [4]	4	(a) visualization of PW strength, (b,c) colored feedback bar, (d) estimated crack time	a) radar chart visualizing PW strength "risks" (e.g. related dictionary words) to remove from the chart, PW information, textual feedback, (b) colored feedback bar, textual feedback, (c) colored feedback bar, textual feedback, information on requirements and impact of additions on PW score, (d) textual estimated crack time, PW information, suggestions	artificial context, interview	within-subject	N = 20 university students and staff	/	other experimental conditions	/	suitability as educational tool: (a) and (c) were found most educative, (b) least educative, (c) was rated best as PW strength indicator
Becker *et al.* (2018) [8]	1	PW lifetime linked to PW strength	colored feedback bar, textual PW lifetime	real university account, field study	pre-post	ca. N = 315,000 PW changes and resets of university accounts	Shannon entropy, user perceptions	pre-post comparison	increase of PW strength on consecutive PW changes and resets	strong passwords were more often reset, policy took 100 days to gain traction, positive reception of PW meter by users
Ciampa (2013) [21]	4	(a, b, c) colored feedback bar, (d) estimated crack time	(a) colored feedback dial, textual feedback, (b) colored feedback bar, textual feedback, (c) colored feedback bar, textual feedback, information on requirements and impact of additions on PW score, (d) textual estimated crack time	artificial context	within-subject, pre-post	N = 66 university students in 4 conditions	entropy based on NIST publication 800-63 R1, score [14]	pre-post comparison, PW change-no change comparison	significantly higher entropy for changed as compared to not changed PWs, descriptive increase in pre-post comparison of changed PWs for (a) - (d)	PW meters (a) - (d) encouraged people with lower ratings to change PW, (d) rated as providing best information but also hardest to understand
Dupuis & Khan (2018) [24]	2	(a) colored feedback bar, (b) social comparison	(a) colored feedback bar, textual feedback, (b) graphical and textual social comparison (colored row of people as strength bar)	artificial context, laboratory study	between-subjects	N = 48 university students in 2 conditions	PW score based on jQuery PW Strength Meter for Twitter	control group (a) with colored feedback bar	significantly higher score for (b) when explicit instruction to create new PW	/
Eargle *et al.* (2015) [25]	5	motivational statements	(a) colored feedback bar and PW information plus motivational statements based on (b) fear, (c) humor, (d) benefits, (e) combined strategies	artificial context, online study	between-subjects	N = 327 Amazon MTurk users in 13 conditions	estimated guess number based on ocl-Hashcat [a], user perceptions	control group with (a) no motivational statement and other experimental conditions	no difference in perceived impact of motivational statement on PW creation, no differences in guess numbers*	mixed perceptions in terms of usability
Egelman *et al.* (2013) [27], Sotirkopoulos *et al.* (2011) [99]	3	(a,c) colored feedback bar, (b) social comparison	Study 1: (a) colored feedback bar, textual feedback, (b) graphical social comparison bar, Study 2: reuse of (a) and (b), (c) vertical colored feedback bar, textual feedback, (d) vertical colored bar, (e) textual feedback, (f) horizontal colored bar, (g) textual social comparison	Study 1: real university account, laboratory study, Study 2: artificial context, online study	Study 1: between-subjects and pre-post, Study 2: between-subjects	Study 1:N = 47 university students and staff in 3 conditions, Study 2: N = 541 Amazon MTurk users in 4 conditions ((d) -(g) dropped)	zero-order entropy, length, composition	control group without PW meter and other experimental conditions	Study 1: significantly higher entropy for (a,b) compared to control, significant pre-post comparison for (a) and (b), not significant between (a) and (b) in entropy but in length, Study 2: no significant difference between (a), (b), (c) and control (conditions (d) - (g) dropped after pre-study)	Study 1: no indication for difference in memorability, Study 2: no significant difference in memorability, longest creation time for (b)
Furnell *et al.* (2018, 2019) [34, 32]	4	(a) colored feedback bar, (b,c) colored emoticons, (d) gamification	Study 1: (a) colored feedback bar, textual feedback, PW information, (b) emoticons as strength indicators, PW information, (c) emoticons as strength indicators, emotive feedback, PW information Study 2: (d) game suggesting characters and scoring PW under time pressure, colored feedback bar	Study 1: artificial context, Study 2: field study	Study 1: between-subject, Study 2: pre-post	Study 1: N = 300 in 5 conditions, Study 2: N = 50	Study 1: weak, medium, strong rating score [68], Study 2: 0 to 100 point rating score, user perceptions	Study 1: control group without PW meter, with only PW information and other experimental conditions, Study 2: pre-post comparison	Study 1: more weak PWs in control, positive impact of any additional PW meter (a)-(c), no significance test, Study 2: stronger PWs after game use, increase in security attitude/ awareness, no significance test	Study 2: positive reception of game by users

Table 5. Results of the literature review 1/5, sorted alphabetically for first author, *sample sizes deemed too small to find significant differences by authors

[a] https://hashcat.net/

Author	No of PWMs	Type of PWM	Design Elements	Setting	Study Design	Participants	Security Measures	Comparison	PW strength	Other outcomes
Furnell & Bär (2013) [33]	1	colored feedback bar	colored feedback bar, textual feedback, PW information	artificial context	between-subject	*N* = 27 university students in 2 conditions	0 to 5 point rating score based on five criteria such as "min. 8 characters", composition	control group without PW meter	significantly stronger PWs	/
Golla *et al.* (2018) [38]	4	(a) colored feedback bar, (b,c) gamification, (d) social comparison	(a) colored feedback bar, textual feedback, PW information, (b) PW high score ranking, PW information, (c) badges from video game, PW information, (d) graphical social comparison bar, PW information	appearance of real account, field study setting	between-subjects	*N* = 302 (mainly university students) in 5 conditions	log-transformed zxcvbn guess numbers [122], user perceptions	control group without PW meter and other experimental conditions	no significant differences	highest preference and longest PW creation time with (b), only 4% clicked PW information, no indication for difference in memorability
Gulenko (2014) [41]	2	affective feedback	feedback bar, PW information, and (a) positive emoticon and message or (b) negative emoticon and message*	artificial context, online study	between-subjects	*N* = 22 in 2 conditions	number of words in passphrase	other experimental condition	significantly more words for (a) compared to (b)	/
Khern-amnuai *et al.* (2017, 2019) [54, 53], Furnell *et al.* (2018) [34]	4	(a) textual feedback and suggestions, (b) estimated crack time, (c, d) social comparison	(a) textual feedback, PW information, (b) textual estimated crack time, PW information, (c) textual social comparison, PW information, (d) textual estimation of PW similarity, PW information	Study 1: artificial context, online study, Study 2: real website account, field study	between-subjects	Study 1: *N* = 500 Amazon MTurk users, Study 2: *N* = 310 website users in 4 conditions	Study 1 and 2: weak, medium, strong score based on Back-off Markov Model by Ma *et al.* (2014) [62]	Study 1 and 2: control group (a) with textual feedback	Study 1 and 2: significantly stronger PWs in (c) compared to (a), people in (c) more often decide to edit PW compared to (a), no significant difference between (a) and (b) or (d)	Study 1: no significant differences in number of clicks on PW information
Kim *et al.* (2014) [55]	1	PW feedback based on similarity and sensitivity across accounts	PW score across all accounts, colored feedback structure for individual PWs, grouping of similar PWs, PW information	artificial context, laboratory study	between-subject	*N* = 48 university students in 3 conditions	PW score from 0 to 100, PW similarity/uniqueness	control group without PW meter and Microsoft PW Checker	significantly higher score and dissimilarity/ uniqueness compared to controls	no significant difference in memorability and perceived usefulness, significantly higher scores for likeliness to use
Komanduri *et al.* (2014) [56]	2	PW guessability	prediction of next character of user's PW before user types it, colored "feedback bar" consisting of number of (un)predicted characters with (a) PW visible and (b) PW hidden per default	artificial context, online study	between-subjects	*N* = 2,560 Amazon MTurk users in six conditions	zxcvbn entropy [122], estimated guess numbers based on PW meter and Kelley *et al.* (2012) [52], PW scores, composition	weakest passwords of all experimental conditions using different password policies	weak PWs were more effectively prevented in (a) and (b) compared to regular PW policies	no significant difference in memorability, higher values in terms of insight, but also annoyance and difficulty
Kulkarni (2010) [59]	1	textual feedback and suggestions	textual feedback, PW information, "pareto-efficient" PW suggestions (based on suggestions and user memorability ratings of suggestions)	artificial context, prototype demonstration	within-subject	*N* = 100 in 4 conditions	score ranging from 1 to 10	control groups Microsoft PW Checker, Google Accounts, and PW policy	increase in descriptive values compared to controls, no significance test	higher descriptive values in terms of memorability, usability and time for PW creation compared to controls, no significance test
Ohyama & Kanaoka (2015) [70]	6	(a,f) colored feedback bar, (b-d) social comparison, (e) motivation	(a) colored feedback bar, (b) two feedback bars for own and similar user's PW, (c) two feedback bars for own and average PW, (d) score feedback for own and similar user's PW, (e) feedback bar with running man, (f) tachometer	artificial context	between-subject	*N* = 600 or 700** Lancers users in 6 conditions	score ranging from 0 to 100 points based on Ur *et al.* (2012) [111]	control group (a) with colored feedback bar	significantly higher score in pairwise comparisons (b) -(f) compared to (a)	/

Table 6. Results of the literature review continued 2/5, sorted alphabetically for first author, * procedure not entirely clear from the description, **conflicting information: *N* = 700, but *N* = 100 in each of the six conditions

Author	No of PWMs	Type of PWM	Design Elements	Setting	Study Design	Participants	Security Measures	Comparison	PW strength	Other outcomes
Ophoff & Dietz (2019) [71]	2	(a) colored feedback bar, (b) gamification	(a) colored feedback bar, textual feedback, (b) overall gamification points score, change in points through PW adjustment	artificial context, online study	between-subjects	*N* = 232 university students and staff in 2 conditions	zxcvbn-score ranging from 0 to 4, estimated guess numbers, crack time [122]	control group (a) with colored feedback bar	no significant differences in terms of score, significantly higher guess numbers for (b) compared to (a)	participants spent more time with (b) gamification points
Peer *et al.* (2020) [76]	6	(a) static textual feedback, (b) colored feedback bar, (c) estimated crack time, (d) social comparison, (e) passphrase suggestion, (f) insertion suggestion	(a) textual feedback, (b) colored feedback bar and textual feedback, (c) estimated crack time, (d) social comparison, (e) passphrase suggestion, (f) insertion suggestion	artificial context, online study	between-subjects and pre-post	*N* = 1,842 Amazon MTurk users in 6 conditions	log-transformed guess numbers based on Melicher *et al.* (2016) [67]	control group (a) with textual feedback, pre-post comparison	significantly higher guess numbers for (b)-(f) compared to (a), significant increase between before and after PW change with PW meter ((a) - (f))	significant moderation effects of user traits on PW meter conditions (except (f) insertion suggestion), no significant difference in memorability
Renaud & Zimmermann (2017a, 2017b, 2019) [79, 80, 84]	2	PW compared to image with PW strength distribution	image with PW strength distributions of (a) university or (b) school peers, arrow indicating current PW strength in relation to PW distribution in image	real university account, field study	between-subject	*N* = 497 university students in 5 conditions	zxcvbn-score ranging from 0 to 4 [122], length	control group without PW meter	no significant differences	/
Renaud & Zimmermann (2018, 2019) [82, 84]	1	PW lifetime linked to PW strength	textual feedback on PW lifetime, image of a dachshund with association "the longer, the stronger"	real university account, field study	between-subjects and pre-post	*N* = 672 university students	zxcvbn-score ranging from 0 to 4 [122], length	previous study's control and experimental groups, pre-post comparison	significantly higher score compared to controls and other conditions, significant increase between pre-post comparison	higher number of forgotten PWs as compared to previous study's conditions
Rodwald (2019) [85]	1	estimated crack time and changing graphical scheme	choice of graphic theme (e.g. low-end to high-end car), PW change form, PW information, textual feedback, fear appeal	group 1 artificial context, group 2 real accounts, online study	within-subject	group 1 *N* = 116 and group 2 *N* = 50 university students	length, composition, entropy, user perceptions	control group without PW meter	increase in descriptive values, no significance test	no indication for difference in memorability
Segreti *et al.* (2017) [89]	8	PW suggestions for blacklisted PWs using common sequences	PW suggestion varying in (a,b) no, one or three suggestions, (c,d) character vs. character class suggestions, (e,f) insertion vs. substitution suggestions, and (g,h) resistance to shoulder-surfing	artificial context, online study	between-subjects	N = 1,799 Amazon MTurk users in 12 conditions	length, composition, PW guessability as calculated using the PW Guessability Service (PGS) [114]	other experimental conditions	no significant difference in terms of (a) the number of suggestions, (b) character vs. character class suggestions, (c) insertion vs. substitution suggestions, and (d) resistance to shoulder-surfing	collection of usability measures, no significant difference in memorability
Seitz *et al.* (2016) [91]	4	PW suggestions using decoy effect	(a) colored feedback bar, textual feedback, (b) colored feedback bar, 1 PW alternative (passphrase), textual feedback, (c) colored feedback bar, 1 PW alternative (complex word), textual feedback, (d) colored feedback bar, textual feedback, 2 PW alternatives to own PW that serves as competitor using decoy effect	artificial context, online study	between-subjects	*N* = 83 Prolific users in 4 conditions	length, composition, zxcvbn-score ranging from 0 to 4, estimated guess number [122]	other experimental conditions	significantly stronger PW in (b) compared to (a), no significant differences between (c) or (d) and (a)	no significant difference in memorability across groups, but when a password alternative was accepted, no difference in usability measures across groups

Table 7. Results of the literature review continued 3/5, sorted alphabetically for first author

Author	No of PWMs	Type of PWM	Design Elements	Setting	Study Design	Participants	Security Measures	Comparison	PW strength	Other outcomes
Shay *et al.* (2015) [93]	3	textual feedback and suggestions for PWs not meeting strength requirements	(a) textual feedback for requirements list, (b) textual feedback for requirements list, character insertion instruction, (c) textual feedback for requirements list, character insertion suggestion	artificial context, online study	between-subject	N = 6,435 Amazon MTurk users in 9 conditions	estimated guess number based on Weir *et al.* (2009) [121], length, composition, user perceptions	control conditions with only PW policy and other experimental conditions	no significant difference between (a) and control, (a) significantly stronger than (b, c), requirement feedback in (a) increased perception but not actual strength compared to control	no significant difference in memorability, (b) had longest creation time
Shepherd & Archibald (2017) [95], Shepherd, Archibald & Ferguson (2017) [96]	4	affective feedback	browser extension detecting risky behavior and providing (a) textual affective feedback, (b) textual and avatar-based affective feedback, (c) textual and colour-based affective feedback, (d) textual, colour-based, and avatar-based affective feedback	artificial context, laboratory study	between-subjects	N = 72 university students and staff in 5 conditions	user perceptions	control group without feedback	people reported not using common PWs, but log-files often revealed common elements	affective feedback (a-d) was rated to increase security awareness and to encourage learning, but did not make people consider changing their PW
Song *et al.* (2015) [98]	1	colored feedback bar and textual feedback	colored feedback bar and textual feedback for Android Pattern Unlock	real accounts, field study	between-subjects	N = 101 EnCloud App users	score ranging from 0 to 1, guess number based on N-gram Markov Model, length, intersecting points, non-repeated segments	control group without PW meter	significantly higher score, length, and intersecting points compared to control, no significant differences in terms of non-repeated segments	/
Stainbrook & Caporusso (2019) [100]	3	(a) colored feedback bar and textual feedback, (b) estimated crack time, (c) social comparison	(a) colored feedback bar and textual feedback, (b) textual estimated crack time, color-coding, textual feedback, (c) graphical and textual social comparison (colored row of people as strength bar)	artificial context	within-subject	N = 115 in 6 conditions	entropy, length	other experimental conditions, PW information, PW policy, and system-generated PW	(a-c) had longest PWs, (a-c) and system-generated PWs had highest entropy, no significance test	system-generated PWs had shortest sign-in and longest log-in time
Sugai *et al.* (2019) [103]	2	(a) colored feedback bar and textual feedback, (b) gamification	(a) colored feedback bar and textual feedback, (b) password score	artificial context	between-subjects	N = 482 Lancers users in 2 conditions	/	other experimental group	/	no significant difference in usability
Sun, Wang & Zheng (2014) [104]	2	(a) colored feedback bar and textual feedback, (b) gamification	(a) colored feedback bar and textual feedback, (b) colored feedback bar, colored feedback bar with percentage filled for Android Pattern Unlock	artificial context, laboratory study	between-subject	N = 71 university students in 3 conditions	pattern length, number of dots, intersections and overlaps	control group without PW meter and other experimental condition	(a, b) created significantly stronger and longer patterns with more dots and intersections compared to control, no significant difference between (a) and (b)	PW meter in (a, b) was rated as helpful, indication for decrease in memorability
Ur *et al.* (2017) [109], Melicher *et al.* (2017) [66], Habib *et al.* (2017) [42]	5	colored feedback bar and textual feedback	(a) colored feedback bar, (b) textual feedback and PW information, (c) colored feedback bar, textual feedback, PW information, (d) colored feedback bar, textual shoulder-surfing resistant feedback, PW information, (e) colored feedback bar, textual feedback, PW suggestion, and PW information	artificial context, online study	between-subject	Study 1: N = 2,717 and Study 2:N = 1,792 Amazon MTurk users in 18 and 8 conditions	length, PW guessability as calculated using the PW Guessability Service (PGS) [114], composition	control group without PW meter and other experimental conditions	for 1class8 PWs: significantly higher guess number for (a)-(e) compared to control, significantly higher guess number for (b) - (e) compared to (a)	positive user reaction to PW feedback, but increase time to create PW, perceived annoyance, and difficulty, no significant difference in memorability

Table 8. Results of the literature review continued 4/5, sorted alphabetically for first author

Author	No of PWMs	Type of PWM	Design Elements	Setting	Study Design	Participants	Security Measures	Comparison	PW strength	Other outcomes
Ur *et al.* (2012a, 2012b) [111, 112]	14	colored feedback bar and textual feedback	(a) textual feedback, PW information, (b) colored feedback bar, textual feedback, (c) colored feedback bar, textual feedback, PW information, variations of (c): (d-g) in appearance of bar, (h-k) in scoring, (l-m) in scoring and appearance of bar, (n) dancing bugs bunny as strength indicator	artificial context, online study	between-subject	N = 2,931 Amazon MTurk users in 15 conditions	estimated guess numbers based on Kelley *et al.* (2012) [52], length, composition	control group without PW meter and other experimental conditions	all conditions except (a) significantly longer compared to control, two variations of scoring (h-k) significantly stronger compared to control	no significant difference in memorability, longer PW creation time and editing with PW meters compared to no meter
Vance *et al.* (2013) [116]	2	(a) colored feedback bar and textual feedback, (b) estimated crack time	(a) colored feedback bar, textual feedback, (b) textual estimated crack time, fear appeal, PW information, graphical elements and symbols	real account, field study	between-subjects	N = 354 international socwall.com users in 4 conditions	estimated crack time	control groups without PW meter, only with PW information, and other experimental condition	significantly longer crack time for (b) compared to (a) and controls, no significant difference between (a) and controls	/
Woo & Mirkovic (2018) [125]	3	colored feedback bar and textual feedback	(a) colored feedback bar, textual feedback and PW information, (b) colored feedback bar, textual feedback and PW information based on Ur et al. (2017) [109], (c) colored feedback bar and textual feedback	artificial context, online study	between-subjects	N = 1,438 Amazon MTurk users in 7 conditions	length, estimated guess number based on Dell' Amico & Filippone (2015) [23]	control groups with PW policy and other experimental conditions	longest PWs in (a) and (c), higher guess numbers in (a) compared to (b), no significance test	highest recall rates in (a), longer PW creation times for (a-c) compared to control groups, no significance test
Xu & Han (2019) [126]	1	textual feedback	textual feedback, PW information	artificial context, PW meter demonstration	between-subjects	N = 50 university students	user perceptions	no comparison	PW meter perceived as increasing security awareness (70%) and encouraging PW change (58%)	/

Table 9. Results of the literature review continued 5/5, sorted alphabetically for first author

Evaluation of the affective reaction

Creating a password with the provided password feedback I felt...

- helpless - capable
- uncertain - certain
- incompetent - competent
- bad - good
- bored - excited
- anxious - assured
- vulnerable - protected
- criticized - appreciated
- ashamed - proud

Evaluation of the features of password feedback

- Items to evaluate the overall feedback, the text feedback, the password suggestions, and the visualization using the items based on Ur *et al.* [109]

Fear Appeal Variables

- items of the fear and self-assurance sub scales of the PANAS-X questionnaire by Watson *et al.* [119]
- items for perceived threat, perceived susceptibility and perceived severity based on questionnaires developed within the framework of the Technology Threat Avoidance model by Liang and Xue [60] and variations of these items as described by Arachchilage and Love [6] as well as Gerber *et al.* [35]

Control Variables

- Affinity for Technology Interaction (ATI) scale by Franke *et al.* [30]
- Password creation sub scale of the security behavior intentions scale (SeBis) by Egelman and Peer [26]
- Self-Efficacy scale (ASKU) by Beierlein *et al.* [9]
- The six items with the highest factor loadings of the Social Norm Espousal Scale by Bizer *et al.* [12]
- Self-constructed social norm compliance items (answered on a 5-point scale ranging from doesn't apply at all to applies completely) based on a pilot study by (blinded for review):
  - The opinion of others is important to me.
  - I often wonder what other people might think of me.
  - The opinion of other people does not influence my behavior.
  - I usually adapt my behavior to fit the norm.
  - I do not care what other people think of me.

Demographic Information

- Age
- Gender
- Education
- Occupation
- IT Security Background

Appendix D: Descriptive Values

Experimental Condition	N	Password Strength			Password Length			Password Entropy		
		M	SD	Md	M	SD	Md	M	SD	Md
Original/Control										
Simple Nudge	59	50.96	28.62	49.48	12.80	6.33	11.00	74.74	35.81	66.73
Information	62	53.58	29.11	59.92	12.23	3.99	12.00	73.56	28.43	71.45
Hybrid Nudge	56	59.94	29.58	64.32	13.93	4.94	13.5	83.90	32.49	83.36
The Person										
Fear Appeal Nudge	58	66.07	27.50	71.48	14.55	5.30	15.00	87.55	31.19	89.31
Motivation Nudge	56	74.45	29.64	81.39	17.04	7.04	16.50	103.20	42.49	105.23
The Password Creation Context										
Compensation Nudge	57	61.14	32.07	69.59	14.37	5.67	14.00	85.44	38.63	89.31
Reciprocity Nudge	56	66.43	31.15	71.50	15.00	5.56	15.00	90.78	36.68	89.31
The Social Context										
Descriptive Norm Nudge	57	68.61	24.60	71.89	14.95	4.27	15.00	90.35	28.34	89.31
Injunctive Norm Nudge	59	58.11	32.40	67.09	14.07	5.63	14.00	85.24	36.64	85.99

Table 10. Descriptive password strength, length and entropy values.