Augmenting Password Strength Meter Design using the Elaboration Likelihood Model: Evidence from Randomized Experiments

Warut Khern-am-nuai
Desautels Faculty of Management, McGill University
1001 Sherbrooke St W, Montreal, QC H3A 1G5, Canada
warut.khern-am-nuai@mcgill.ca

Matthew J. Hashim
Eller College of Management, University of Arizona
1130 E Helen St, Tucson, AZ 85721, United States
mhashim@arizona.edu

Alain Pinsonneault
Desautels Faculty of Management, McGill University
1001 Sherbrooke St W, Montreal, QC H3A 1G5, Canada
alain.pinsonneault@mcgill.ca

Weining Yang
ByteDance Inc.
250 Bryant St, Mountain View, CA, 94041, United States
weining.yang@bytedance.com

Ninghui Li
Department of Computer Science, Purdue University
305 N University St, West Lafayette, IN 47907, United States
ninghui@purdue.edu

Abstract

Password-based authentication is the most commonly used method for gaining access to secured systems. Unfortunately, empirical evidence highlights the fact that most passwords are significantly weak and encouraging users to create stronger passwords is a significant challenge. In this research, we propose a theoretically augmented password strength meter design that is guided by the Elaboration Likelihood Model of persuasion (ELM). We evaluate our design by leveraging three independent and complementary methods: a survey-based experiment using students to evaluate the saliency of our conceptual design (proof-of-concept), a controlled laboratory experiment conducted on Amazon MTurk to test the effectiveness of the proposed design (proof-of-value), and a randomized field experiment conducted in collaboration with an online forum in Asia to establish proof-of-use. In each study, we observe the changes in users' behavior in response to our proposed password strength meter. We find that the ELM augmented password strength meter is significantly effective at addressing the challenges of password-based authentication. Users exposed to this strength meter are more likely to change their password, leading to a new password that is significantly stronger. Our findings suggest that the proposed design of augmented password strength meters is an effective method for promoting secure password behavior among end users.

Keywords. password strength meter, design science, elaboration likelihood model, randomized experiment

1. Introduction

Passwords are the dominant authentication mechanism implemented by organizations for gaining access to secured systems (Steinbart et al. 2016). Accordingly, issues related to password security are critically relevant to organizations, and techniques that encourage end-users to exhibit better password-generating behaviors are generally valuable (Furnell et al. 2018). Despite the clear importance of improving password security, some experts assert that password-related issues will be irrelevant in the near future (e.g., Bachmann 2014). However, practical evidence suggests otherwise. For example, even organizations such as the US Department of State—required by law to implement enhanced security—have struggled to move beyond passwords for securing systems (Cimpanu 2018). In the meantime, several techniques designed to minimize password-related issues may generate unexpected consequences. For instance, researchers have found that some popular password managers were affected by critical vulnerabilities that expose user credentials (Wei 2017). As a result, using password management software with complex passwords may not necessarily address all password issues. Meanwhile, the cost of implementing multi-factor authentication technology has increased significantly due to the inadvisable and insecure use of text messages and emails for receiving one-time passwords (Maida 2016). The technology is thus prohibitively expensive for small and medium businesses.

With the unexpected impracticability of alternative techniques to alleviate password security issues, the spotlight on an effective solution has now refocused on persuasive techniques such as enhancing password strength meters (Kariryaa and Schöning 2020; Yang et al. 2020). In this paper, we design an augmented password strength meter with the following characteristics. First, the design is driven by a well-known psychology theory. Second, it significantly enhances users' password-generating behavior without requiring the recently discouraged password complexity policy per the Digital Identity Guidelines NIST SP 800-63 (Grassi et al. 2017). Third, it can be readily implemented in a wide variety of settings with minimal cost.

Password strength meters calculate the complexity of an input password and display the password's strength to users, usually ranging from weak (low complexity; less secure) to strong (high complexity; more secure). Ideally, users who initially choose weaker passwords will receive feedback and then consider revising their chosen password. Although previous literature has shown that password strength meters generally lead

users to create stronger passwords in a controlled setting (Egelman et al. 2013; Shay et al. 2015), the effectiveness of password strength meters in reality continues to be a controversial topic. For one, even with the presence of password strength meters in many websites and systems, the problematic issue of weak passwords persists (Slain 2016; Verizon 2017). As a result, many recent works have aimed to improve the effectiveness of password strength meters. The primary focus of this stream of work is on enhancing the algorithms that are used to calculate password strength with the key thesis that traditional password strength meters are not effective because the strength displayed by the meters are not accurate. In that regard, although it has been generally recognized that password strength meters are less accurate than an entropy measurement of an ideal case (Houshmand and Aggarwal 2012; Weir et al. 2010), many advanced techniques have been proposed to overcome this limitation such as the use of probabilistic context-free grammar (e.g., Veras et al. 2014; Weir et al. 2009) and Markov models (e.g., Castelluccia et al. 2012; Ma et al. 2014). In addition, several alternative metrics to measure password strength have been introduced such as simulating password cracking algorithms (Kelley et al. 2012) and password ranking algorithms (Yang et al. 2016).

In this research, we complement the mainstream research on password strength meters by examining issues surrounding traditional password strength meters from an alternative aspect. Particularly, we study the presentation component of password strength meters when the algorithm to calculate password strength is held constant. Our work is built upon an alarming finding by Furnell (2011) that theoretically grounded guidelines for password strength meter interface design are scarce. Hence, the design and implementation of most password strength meters both in the literature and in practice usually operate in a black-box fashion (i.e., without explanations or justifications of design choices) from the users' perspective (Carnavalet and Mannan 2014), making it difficult for users to understand (Carnavalet and Mannan 2015). Inspired by such an ongoing gap in the literature, this paper identifies a theory from the psychology literature that underpins the core function of password strength meters. Specifically, we draw upon the Elaboration Likelihood Model of Persuasion (ELM; Petty and Cacioppo 1986) to develop a theory-guided password strength meter design as suggested in design science literature (Gregor and Hevner 2013). Following that, we utilize multiple randomized experiments to establish the proof-of-concept, proof-of-value, and proof-of-use of our theoretically augmented

design of password strength meters. We find that the proposed password strength meter enhances password security as theorized. It effectively nudges users to change their passwords more readily, and users also choose new passwords that are significantly stronger than the passwords initially entered into the system.

Our research contributes to theory and practice. In terms of theoretical contributions, we are among the first to establish a theoretical foundation of how password strength meters interact with and influence users. We then utilize such a foundation to propose an augmented password strength meter that can improve users' password-generating behavior. We demonstrate through multiple experiments that our improved design is a promising approach for promoting secure behavior among end users. As regards contributions to practice, a novelty of our proposed design lies in the fact that the implementation requires minimal changes to both the server- and the client-side. That is, our design does not require significant computing resources on the server-side, thus an investment in new hardware is not necessary, unlike other password security technologies such as authentication manager software. Further, the modifications required to existing password strength meter interfaces are simple and relatively straightforward from a client-side software perspective. Our method also works well in stimulating users to improve their secure behavior for both higher-security systems (e.g., online banking), and lower-security systems (e.g., online forums).

The rest of the paper is organized as follows. In Section 2, we describe the conceptual design of our password strength meter. We first identify a theory from the psychology literature that underpins the psychological mechanism of password strength meters on users' password-generating behavior. Then, we leverage such an underpinning theory to conceptually design a theoretically augmented password strength meter. Subsequently, we describe the implementation of our conceptual design and propose the evaluation method in Section 3. Following that, we present the survey-based experiment and related results in Section 4, followed by the laboratory experiment in Section 5, and the field experiment in Section 6. Finally, we discuss the implications of our results and conclude our study in Section 7.

2. Theoretically Augmented Design of Password Strength Meters

2.1 Background Literature

In this section, we discuss prior studies that are closely related to our research objective to examine the presentation component of password strength meters. For a comprehensive literature review on password strength meters in general, see Golla and Dürmuth (2018).

Among existing works in this area, Ur et al. (2012) propose that different visual cues such as the color and size of a progress bar would significantly affect users' perception of the password strength meter's message. Relatedly, Golla et al. (2018) demonstrate that other types of visual cues such as radars, tachometers, and dancing bunnies also work well in attracting users attention to the password strength meter's message. Furthermore, Furnell et al. (2018) propose the inclusion of emoticons to improve users' attention to the password strength meter's message. These prior works consistently demonstrate that their approaches of using visually driven stimuli are significantly effective in improving users' password-generating behavior. However, our understanding on underlying mechanisms that drive the change in users' password-generating behavior after being exposed to visual stimuli, and underpinning theories that can explain such a behavioral change is lacking in the literature (Carnavalet and Mannan 2015). In this study, we respond to a call for a theoretical investigation on users' secure behavior (Karjalainen et al. 2019) by investigating password strength meters using a theoretical lens to bridge this knowledge gap in the literature.

2.2. Elaboration Likelihood Model of Persuasion

At the outset, a password strength meter is a tool that is expected to persuade users to change their password-generating behavior. As such, we establish the theoretical foundations of password strength meters by relying upon the Elaboration Likelihood Model of Persuasion or ELM (Petty and Cacioppo 1986), as ELM is widely used to study the effectiveness of persuasive methods on changing behavior.

ELM is a dual process theory that describes how persuasion leads to changes of attitudes (and hence a change of behavior). It explains how individuals process stimuli through two major routes that lead to persuasion: the peripheral route and the central route. Peripheral route processing involves a low level of

elaboration where the audience is influenced by positive or negative cues in the stimuli or relies upon inferences of the information presented (rather than the content itself). This process does not require much cognition and attention. As a result, individuals who are persuaded via the peripheral route processing usually have a change in attitude and behavior that tends to be superficial and temporary. Meanwhile, the central route processing involves a high level of elaboration where the audience carefully and thoughtfully considers the content of the information presented. This process generally requires a great amount of cognition and attention. As a result, if individuals are persuaded under central route processing, the resulting attitude and behavioral changes tend to be more salient.

Through the lens of ELM, visually driven stimuli used by password strength meter improvements discussed in Section 2.1 seem to be processed through the peripheral route (Reinhard and Sporer 2010). Although these stimuli (e.g., color cues, progress bars) may be effective in nudging users to change their behavior, prior works have consistently shown that peripheral-route-driven change tends to be frivolous and short-lived (Tang et al. 2012). Such a theoretical guideline may help explain why the issue of weak passwords continues to persist even though password strength meters with visual cues are already widely implemented. Meanwhile, prior IS literature has demonstrated that the use of elaborate message that stimulates users' cognitive thought, which tends to be processed through the central route according to ELM, works significantly well in persuading users to change their behavior saliently in the context such as technology acceptance (Bhattacherjee and Sanford 2006) and adoption of electronic health records (Angst and Agarwal 2009). However, there are virtually no prior works that augment password strength meters using a stimulus that is likely to be processed through the central route. Inspired by this gap in the literature and practice, we design, develop, and validate an augmented design of password strength meters that aims to persuade users via the central processing route according to ELM. We present our conceptual design in the next subsection.

2.3 Conceptual Design

We follow the design science literature to conceptually design an augmented password strength meter using the theoretical driven design principle. We start by presenting a type V theory defined as "design and action" in Gregor (2006). In essence, this type of theory, which is essential in developing actionable research insights,

provides explicit prescriptions for constructing an artifact. Recall that ELM suggests that a persuasion tends to be processed through the central route when it invokes the audience's cognitive curiosity (e.g., involvement, urgency, interest). As such, the resulting change in behavior tends to be more salient (Tang et al. 2012). Therefore, the key prescriptive design element here is to identify and incorporate stimuli into password strength meters that are likely to be processed through the central route (e.g., messages that are likely to stimulate users' cognitive thought). For that, we survey prior literature that utilizes ELM to propose or explain the effectiveness of persuasive messages in other empirical contexts.

The first type of persuasive messages that are commonly used in prior ELM-related studies is fear appeals as prior literature has shown that fear appeals messages tend to instigate cognitive awareness (Ruiter et al. 2001). For example, Munoz et al. (2010) propose that the use of fear appeals messages are effective in promoting responsible gambling behavior. Fear appeals messages are also used in other contexts such as reducing risk of contagious diseases (Dinoff and Kowalski 1999). Relatedly, fear appeals is also commonly used to motivate users in the context of information security (Jenkins et al. 2014; Vance et al. 2013), which influences user behavior both in individual settings (Johnston and Warkentin 2010) and in organizational settings (Boss et al. 2015; Herath and Rao 2009). Therefore, we propose **fear appeals** as the first type of persuasive message that could be incorporated to a password strength meter to improve password-generating behavior.

Second, several previous works use ELM to explain the effectiveness of peer pressures. The core concept here is that peer pressure tends to play an important role in invoking cognitive thinking (Alegre Ansuategui and Moliner Miravet 2017), especially for an indirect type of peer pressure called peer comparisons (Festinger 1954; regarding reliance on social comparisons). Further, Petty et al. (1991) demonstrate how peer comparisons messages can be used to persuade individuals in the context of drug abuse prevention. This type of message has also proven effective at motivating individuals in group trainings (Ernst and Heesacker 1993). In the context of information security, peer comparisons are also commonly used in several related context including privacy management systems (Goecks and Mynatt 2005) and security awareness training (Ophoff and Janowski 2015). Therefore, we propose **peer comparisons** as the second type of persuasive message in this study.

Third, past literature utilizes ELM to theorize the positive influence of a common bond in motivating or persuading individuals. That is, messages related to common bond tend to induce cognitive processes (Postmes and Spears 2000). For one, Cyr et al. (2018) uses ELM to investigate the persuasiveness of common bond in website designs. In the same vein, Chung and Han (2017) demonstrate the relationship of common bond messages and tourists' persuasion in social media. In the context of information security, especially in password security, common bond messages have a strong potential to impact the audience's cognition because it is widely recognized that commonly used passwords have a higher risk of being guessed in comparison to less commonly used passwords (Pinkas and Sander 2002). As a result, we propose **common bond** as the third type of persuasive message in this study.

Next, we transform our conceptual design into an actual implementation of an augmented password strength meter to develop a "situated implementation of an artifact" (as defined by Gregor and Hevner 2013). We will then evaluate the performance of this design through a series of experiments where we connect the outcomes of these experiments back to our design principle and ELM. The first experiment serves as a proof-of-concept where we use a survey-based experiment to gauge whether the audience of our augmented password strength meter carefully considers the messages (i.e., whether our messages are likely to be processed through the central route as intended). Following that, a proof-of-value experiment is conducted in a controlled laboratory setting to examine whether our augmented password strength meters lead users to change their behavior. Lastly, we generalize our results by performing a proof-of-use experiment where our augmented password strength meters are implemented in the field. We describe how our conceptual design is specifically implemented and evaluated in the next section.

3. Design Implementation and Evaluation Method

In this section, we describe the implementation of the proposed conceptual design. Specifically, we discuss the baseline password strength meter and how the strength of a password is calculated. Following that, we explain how each type of proposed persuasive messages is integrated into password strength meters. Lastly, we propose evaluation methods that validate the performance of our designs.

3.1 Baseline Password Strength Meter

We first describe the password strength meter that will be used as the baseline in our study, along with how the strength of a password is calculated. Note that password strength can be measured in many different ways, and an industry or academic standard does not exist. In our work, we adopt an approach that calculates password strength by using a probabilistic model called the Backoff Markov Model, which is consistently cited as one of the best models for our type of experiments (e.g., Ma et al. 2014; Ur et al. 2015). Specifically, we calculate password strength in our experiments using the Backoff Markov Model with end symbol normalization proposed by Ma et al. (2014).[1] The model uses machine learning to predict the probability that a given password will be chosen. The Backoff Markov Model allows us to maintain consistent comparisons of password strength across experimental treatments. To ensure realism and provide a baseline for password strength calculations, we train the model using the RockYou password dataset.[2] This dataset contains over 32 million passwords in plaintext from the publicly leaked social application site RockYou from December 2009, making it one of the largest plaintext password lists. It is also considered one of the best password lists available by several security researchers (e.g., Bowes 2015). In this paper, we use the RockYou list provided by SkullSecurity, which only contains the passwords while other information (e.g., username, email) is removed.

The baseline meter displays a strength label as shown in Figure 1 (a strong password has been entered in the example shown). Displaying a strength label to the user is a standard practice by many major websites and systems (Furnell 2011). Our meter displays one of the following strength labels based on the results generated by the Backoff Markov Model from the password entered: Weak, Medium, Strong. The password is considered "Weak" if its probability calculated by the model is within the top 300,000 weakest passwords. Specifically, if an attacker tries passwords following a descending order of the probability generated by the model, the account will be compromised in 300,000 attempts. The password is considered "Medium" if its probability is greater

[1] Further details regarding the Backoff Markov Model are available in Appendix A. As a robustness test, we also use an entropy-based measurement (Taha et al. 2013) to calculate password strengths instead. The results remain qualitatively similar with both password strength calculation approaches. Details are available in Appendix E.4.

[2] We also consider the Yahoo dataset (leaked in 2012, contains approximately 450,000 records) as an alternative training dataset. The labels of passwords generated by the alternative model are consistent with those of the main model (approximately 2% of the passwords have different labels between these two training datasets).

than the top 300,000 but less than the top 5,000,000 weakest passwords. Finally, if the probability of a password is not in the top 5,000,000 weakest passwords, we label the password as "Strong". These thresholds are chosen based on guidelines used in previous password strength research (e.g., Ur et al. 2012). Note that these strength thresholds are always the same across all password strength meter types used in our study.

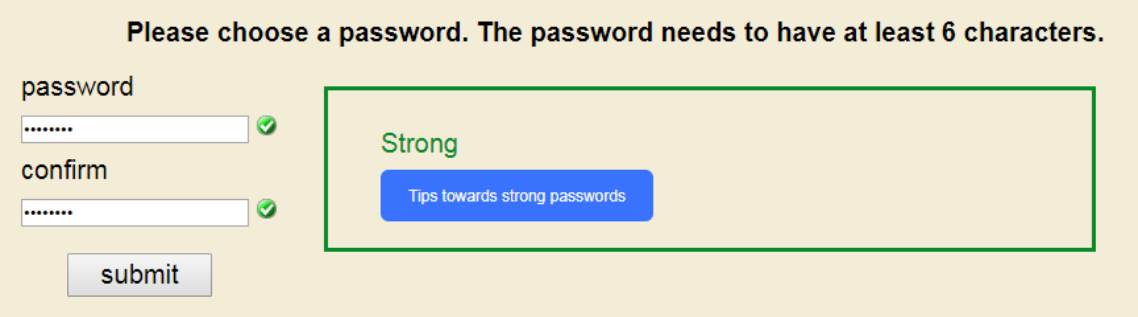

Figure 1: Password Generation Interface

The strength of the entered password (and any additional warning message) is calculated and shown in all treatments when the user moves their cursor away from the input box (i.e., after the user types in the input box, but before the user clicks "submit"). That is, our web-based software listens to an event of the password input field in Javascript called *focusout*, which is triggered when an element loses focus (e.g., when a cursor is out of an input field). Once an event is triggered, the script estimating and displaying the strength of passwords is executed. All necessary data are transmitted to users' browser by using Javascript and AJAX calls when the webpage is loaded, and password strength is calculated based on the Backoff Markov Model. We then derive the persuasive message that will be shown to the user as described in the following subsections.

3.2 Password Strength Meter with Fear Appeals Message

With the use of fear appeals in prior ELM-related studies in other contexts, there are many mechanisms that may be used to manipulate fear appeals in a password security setting. One such mechanism is the perception of susceptibility and vulnerability to a threat. In the case of our proposed password strength meter, we persuade users by incorporating fear appeals into the warning message feedback. A fear appeal is comprised of a message that includes the severity and susceptibility of a threat, and the ability to alleviate the threat (Johnston and Warkentin 2010). Particularly, our design displays the calculated average time that would be required to crack an entered password by estimating the time needed for an attacker to continuously attack the account, starting

from the most popular passwords. Based upon real-world scenarios, we implement a conservative assumption that the attacker can try 100 passwords per second (Mirante and Cappos 2013). Under a scenario where a weak password was entered into the meter, an example of the fear appeals warning message is "Weak. We estimate that it takes 10 seconds to crack your password, assuming that the attacker can try 100 passwords every second." The time displayed to the user invokes the perceived threat susceptibility of not generating a strong password. That is, a short amount of time calculated by the password strength meter would increase the saliency of susceptibility and vulnerability to the threat of the password getting cracked quickly, invoking a fear appeal to generate a stronger password. Figure 2 illustrates an example of the fear appeal message incorporated in a password strength meter under a weak password.

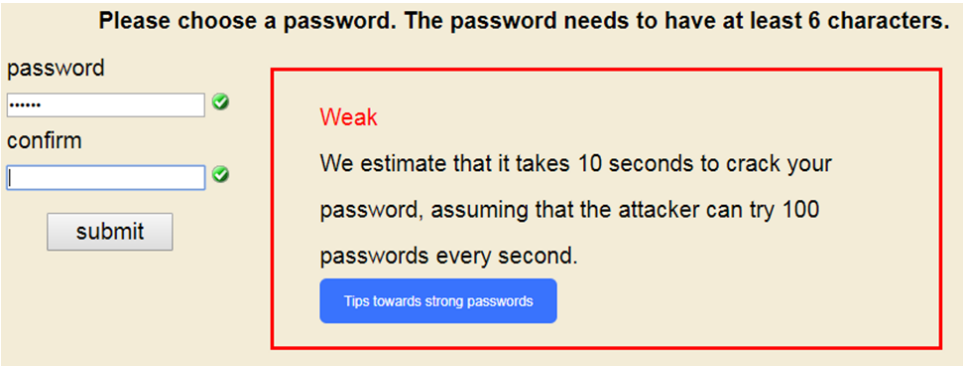

Figure 2: Password Strength Meter with Fear Appeals Message

3.3 Password Strength Meter with Peer Comparison Message

Prior ELM-related works have consistently demonstrated that peer comparison messages tend to be effective in persuading users to saliently change their behavior. Therefore, we consider a message design that stimulates users' password-generating behavior using peer comparisons in this subsection. Peer comparisons can result from the display of peer data (e.g., Bolton 2014; Ophoff and Janowski 2015), and awareness of oneself in group settings. In that vein, a message provides peer data to situate oneself in a peer comparison. The pressure arising from that comparison manifests as the individual taking action to control and/or improve their password based upon how it compares to others. Specifically, we designed the password strength meter to include a warning message that contains the rank of the password in comparison to an implied leaderboard. The rank is calculated by first generating a list of passwords using the Backoff Markov Model described earlier. Following that, the

passwords are sorted into descending order by the probability of the password matching other passwords from the model output. That is, weaker passwords (i.e., those with a higher probability of matching other passwords) are output earlier and with a lower rank, whereas stronger passwords will have a higher rank. The list of ranked passwords is then used as a set to determine the rank of the password entered in the password strength meter by a participant. Specifically, we pick passwords that have a rank of only one significant digit (e.g., 10, 20, ..., 100, 200, ...) from the list and calculate the corresponding probability, resulting in a rank-probability pair for each password. Then, we calculate the probability of the entered password, and match the probability with the closest rank-probability pair generated earlier. The rank of the password is estimated as the highest rank whose corresponding probability is smaller than the probability of the entered password. In this way, we essentially round up the actual rank of the password and display one significant digit. Under a scenario where a weak password was entered into the meter, an example of the peer comparison warning message for a password of rank 291 is "Weak. We estimate that the password you chose is among the 300 weakest passwords." Figure 3 illustrates an example of the peer comparison message incorporated in a password strength meter when a week password was entered.

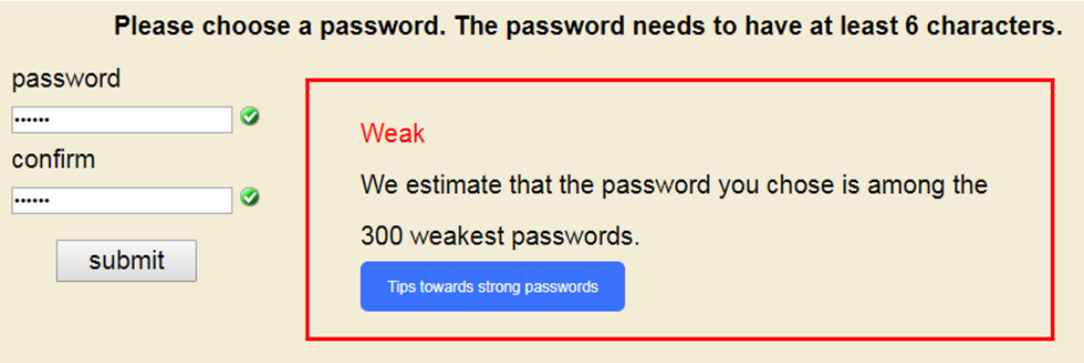

Figure 3: Password Strength Meter with Peer Comparison Message

3.4 Password Strength Meter with Common Bond Message

Messages that incorporate a common bond are particularly effective in persuading users in prior ELM-related studies in other contexts. In addition, common bond is a critical factor during the decision-making process when faced with risk (Rubinstein 1988). In the password meter context, displaying a probability to a user is one of applicable methods for communicating the level of risk (e.g., Slovic et al. 2000; Visschers et al. 2009). That

is, the level of bond between a given password and other passwords can be communicated by calculating the probability that the given password will match those passwords already entered into the system. Therefore, if a password has a strong bond (e.g., high probability of match), then the user will face an increased level of risk of their password being breached by an attacker in comparison with those passwords that have a weak bond.

To persuade users using a common bond message, we designed the password strength meter to include a warning message that contains the calculated strength of the password entered into the password strength meter, along with an estimated number of accounts that use the same password. The estimated value is based upon an assumption of 10 billion accounts globally, using the Backoff Markov Model. An example of the similarity warning message for a password having a probability of matching others in the list of 0.001 is "Weak. We estimate that about 10,000,000 other accounts will have the same password as you within 10 billion accounts." Figure 4 illustrates an example of the common bond message incorporated in a password strength meter when a weak password was entered.

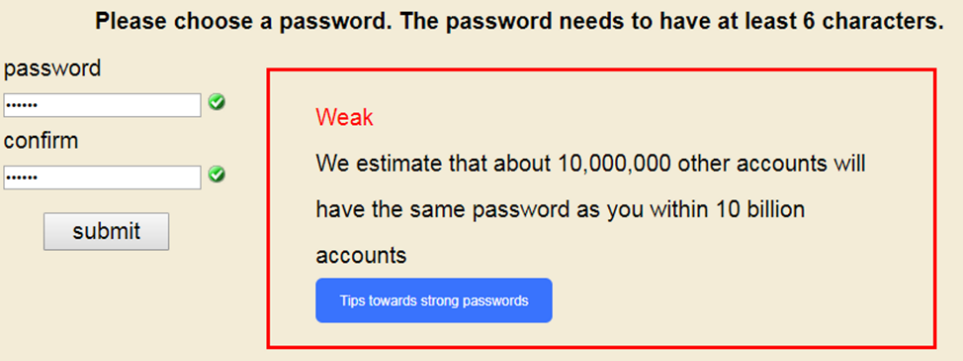

Figure 4: Password Strength Meter with Common Bond Message

3.5 Design Evaluations

Evaluating the performance of our proposed design in the context of password-generating behavior can be challenging because such a behavior is dynamic and complex. Therefore, we utilized a multimethod approach by developing three independent quantitative studies that complement each other. By using multiple methods, we seek to explore password strength meters from multiple views, reinforce findings, and address weaknesses in each independent approach. Accordingly, we first use a survey-based study to understand how the users react to different types of the proposed persuasive message (i.e., whether the messages we provide are likely to be

processed through the central route according to ELM as intended). In addition, the survey allows us to compare if different types of persuasive message make different behavioral constructs salient for respondents. Therefore, the survey can help establish the proof-of-concept of our proposed design. Next, we evaluate the effectiveness of our proposed augmented password strength meter through a randomized controlled experiment. Such an experiment allows us to maximize the internal validity of the design by limiting possible confounding factors and selection effects that may occur in other research methodologies. Hence, the controlled experiment allows us to establish the proof-of-value of our proposed design. Finally, some concerns could be raised regarding external validity due to the artificiality of the setting used in the laboratory experiment. For that reason, we validate the results by partnering with a website to conduct a field experiment. In that regard, the field experiment helps in validating the proof-of-use of our proposed design. In summary, our evaluation methods achieve three purposes of multimethod research: complementarity of views by enhancing the results of persuasive messages, corroboration of results across studies, and compensation by using multiple methods to eliminate the weaknesses of each individual method (Venkatesh et al. 2013). Our design also fulfills the design science paradigm of going the last research mile (Nunamaker et al. 2015).

Unless otherwise noted in each study, we implemented four treatments using a between-subjects design with random assignment for each of the experiments that follow in Sections 4 to Section 6. Three of the treatments make use of the augmented password strength meter with persuasive messages we proposed, whereas the fourth serves as a control treatment. That is, the control treatment only includes a warning message for the label derived from the calculated strength of the password entered (e.g., Weak, Medium, Strong). Meanwhile, the three treatments displayed persuasive messages based upon one of the following: Fear Appeals (*Time*), Peer Comparisons (*Rank*), or Common Bond (*Probability*).

Table 1 provides the link between the type of persuasive message and experimental treatment. Note, a password length of at least six characters was required for all treatments. No other requirements for password complexity were required. Overall, each of the three treatments (time, rank, probability) provides a warning message communicated by the password strength meter to the participants such that the message is likely to provoke the recipients' cognitive thought regarding their input password (i.e., likely to be processed through

the central route according to ELM). Each treatment (including control) always retains the password strength label, and the strength of an entered password will always be the same across all treatments. Therefore, our implementation of the conceptual design of the augmented password strength meter provides a controlled manipulation of the warning message, and an ideal point of comparison for experimentation. In the next three sections, we discuss a set of studies used to evaluate our proposed design and the corresponding results.

Table 1. Operationalization of Persuasive Messages as Experimental Treatments

Message Type	**Treatment**	**Warning Message**
None	*Control*	Strength
Fear Appeals	*Time*	Strength + Time to crack
Peer Comparisons	*Rank*	Strength + Comparison to others'
Common Bond	*Probability*	Strength + Popularity of password

4. Study 1: Evaluating Persuasive Message Saliency

The core principle of our proposed design is to use persuasive messages that are likely to be processed through the central route according to ELM to enhance users' comprehension of password security which, in turn, improves password-generating behavior. However, it is important to note that persuading users through central route processing is not necessarily trivial and improperly designed stimuli usually fail to take the central route (Petty et al. 1983). So, we first seek to determine if our proposed stimuli (i.e., persuasive messages) are successfully processed through the central route by the message recipients (i.e., whether the message recipients carefully and comprehensively consider the messages). Here, we use surveys to measure how users respond to our proposed designs in comparison with a standard password strength meter to demonstrate proof-of-concept for further studies. In other words, do we observe evidence that persuasive messages allow users to carefully and comprehensively consider the implications of having stronger passwords? We implement an exploratory survey-based study to answer that question.

4.1 Survey Design

We draw upon two previous studies used in security research to assist in evaluating the saliency of our persuasive messages. First, Johnston and Warkentin (2010) have shown that compliance with information security recommendations can be influenced by persuasive messages, notably those messages relating to fear

appeals. Their research has been widely cited by the literature and the measures used are easily adaptable to research using password strength meters. We also draw upon Cho et al. (2010) and their work on optimistic bias. We do so because optimistic bias considers how individuals judge their own and others' risk when faced with a decision affecting information security contexts. Constructs from these two theories are also used in conjunction with ELM in the literature to examine the processing route of a stimulus (e.g., Ahn et al. 2014; Dunbar et al. 2014). That is, if individuals respond differently to the warning messages in comparison to a control treatment, then persuasive messages would appear to be processed through the central route processing in ELM as intended and provide initial evidence that users carefully and comprehensively consider the messages, which may eventually enhance their password-generating behavior.

Note that the purpose of this survey-based study is not to develop new theory. Instead, our goal is to explore potential changes in well-known models for measuring information security behaviors, thereby providing evidence of comprehension of the warning messages and motivation for additional follow-up studies. Therefore, we use the constructs defined by Johnston and Warkentin (2010) and Cho et al. (2010), with minor modifications to wording for the password meter context. From the former paper, we utilize Perceived Severity, Perceived Susceptibility, Response Efficacy, and Self Efficacy. These four constructs capture the understanding of the consequences, likelihood of theft, effectiveness, and difficulty of choosing a particular password. From the latter paper, we utilize two measures of Perceived Vulnerability to password theft for a user (Yours) and others in society (Others). These two constructs allow users to situate the risk of their password strength in comparison with other users. The Perceived Vulnerability measures (i.e., Yours, Others) are used by Cho et al. (2010) as separate dependent variables in their model, so we also use them as dependent variables predicted by the two Efficacy constructs from Johnston and Warkentin (2010), which are in turn predicted by Severity and Susceptibility (additional details to follow in Section 4.3).

In Section 2.3, we motivated three types of messages that could be drawn upon for the creation of enhanced warning messages for a password strength meter. To maximize the manipulation of the password strength meter for evaluating message saliency (i.e., display the worst outcome), we use a Weak password warning message as the label shown by the password strength meter in the survey-based experiment.

4.2 Survey Implementation

Surveys were distributed electronically using Qualtrics during fall semester at a large university in the southwest United States. Students were all undergraduates and represented various majors across the management school (e.g., accounting, finance, information systems, marketing, economics, organizational behavior, etc.). Participation was limited to those aged 18 years or older but was otherwise largely homogeneous given we sampled from an undergraduate student population. Participants were made aware that their responses to the survey would be recorded anonymously and confidentially and course credit was awarded to students for survey completion. Attention-catching questions were used throughout the survey to ensure that participants were carefully reading and responding to the survey questions.

Participants were presented with a hypothetical scenario in which they are required to create an account for use on an online discussion forum (similar to Study 3 in Section 6). After reading the scenario, the participants viewed a screenshot of a password strength meter for their randomly assigned treatment. Subsequently, participants were then presented with measures from the constructs described earlier (Cho et al. 2010; Johnston and Warkentin 2010). All survey questions were presented as a Likert-type scale ranging between 1 and 7 such that 1 indicates that the participant strongly disagrees with the statement, and 7 indicates that the participant strongly agrees with the statement. Adjustments for reverse coding were made when necessary to be consistent across measures and are noted in Appendix B. Out of 282 surveys completed, 253 were retained after dropping those responses that were found after further review to be substantially incomplete. Counts of participants assigned to each treatment are as follows: 65 are assigned to the control group, 61 are in time, 64 are in rank, and 63 are in probability. Details regarding randomization checks of participants in this study are provided in Appendix E.5.

4.3 Survey Evaluation

We present the results of the survey using a PLS multi-group analysis to determine if there exists evidence that users carefully and comprehensively consider the added persuasive messages (i.e., likely process the message through the central route processing of ELM), potentially persuading a change in behavior. We conducted all of the preliminary analyses to ensure the reliability of the scales through the appropriate multivariate tests to

establish discriminant validity, convergent validity, and internal consistency (Gefen and Straub 2005). For the sake of brevity, tests for reliability and validity are discussed in detail along with the results of the preliminary analyses in Appendix B.

Table 2 presents the results of a PLS multi-group analysis (MGA). We used SmartPLS 3.2.7 (Ringle et al. 2015) for the analyses and specified bootstrapping with 500 re-samples for calculating the standard errors. The results of our exploratory multi-group analysis in the survey-based study provides evidence that the persuasive messages presented by our augmented password strength meter are salient to participants and manifest as self-reported differences in the measured relationships between commonly used security related constructs in comparison to a control. The results indicate that the users responded differently to the different treatments.

Table 2: Multi-Group Analysis: Control vs. Persuasive Message Treatments

Relationship	**Difference in Coefficient: Control vs. Treated**					
	Time		**Rank**		**Probability**	
Perceived Severity ↔ Response Efficacy	0.16	-	-0.04	-	-0.13	-
Perceived Severity ↔ Self Efficacy	-0.03	-	0.08	-	-0.23	-
Perceived Susceptibility ↔ Response Efficacy	0.37	$p = 0.06$	0.06	-	0.22	-
Perceived Susceptibility ↔ Self Efficacy	-0.26	-	-0.15	-	-0.36	$p = 0.06$
Response Efficacy ↔ Your Perceived Vulnerability	0.24	$p = 0.10$	-0.10	-	0.14	-
Response Efficacy ↔ Others' Perceived Vulnerability	0.22	-	0.40	$p = 0.04$	-0.01	-
Self Efficacy ↔ Your Perceived Vulnerability	-0.21	-	0.04	-	-0.27	$p = 0.06$
Self Efficacy ↔ Others' Perceived Vulnerability	0.04	-	-0.19	-	-0.13	-

We focus our analysis on the individual effectiveness of each persuasive message treatment in comparison to a standard password strength meter (i.e., our control treatment). The fact that there is no overlap of significant results when comparing the treated groups to the control group (Table 2) indicates that the persuasive messages seem to be activating the salience of the messages through different constructs. Further, the constructs that are activated are consistent with the type of messages that were used in each password conditions. For instance, when comparing "Rank" to the control group (center column in Table 2), we observe a significant and quite large difference for the relationship between response efficacy and others' perceived vulnerability. Recall the message conveyed in the rank treatment persuades the user to take action by comparing oneself to a group of others. Accordingly, we observe a significant result that allows individuals to judge that others would be less effective at their ability to respond to the challenges of choosing a strong password. Similar

evidence arises for the time and probability treatments, in that there are two relationships for each treatment that indicate significant differences in coefficients in comparison to the control treatment. Comparing "Time" to the control group (left column in Table 2) we observe significant effects from the persuasive message that heighten the susceptibility of the password being stolen, correlating to difficulty using hard-to-guess passwords and increased vulnerability towards becoming a victim. These two results are consistent with the meaning of the message in the time treatment. In this case, the message suggests the participants password would be cracked in 10 seconds, making salient the perception of a helpless task. Similarly, the comparison of "Probability" to the control group (right column in Table 2) indicates that self efficacy is negatively associated with both susceptibility and their perceived vulnerability. At first glance, we might have expected positive relationships instead of negative. However, upon a closer examination at the meaning of the message, the results are not so surprising. The treatment message suggests the individual faces a probability that can easily be improved (activating self efficacy), and therefore can make the password stronger.

The results of the preliminary proof-of-concept study suggest that individuals appear to process the stimulus through the central route as designed. They tend to understand the implications of the password strength meter messages better and their reactions are consistent with the type of persuasive message of each treatment. Therefore, we can expect that users will likely change their behaviors if the password strength meter is implemented. This is the focus of our next two experiments.

5. Study 2: Evaluating the Design in a Laboratory Experiment

Next, we evaluate the effectiveness of our proposed password strength meter in a randomized controlled experiment to establish proof-of-value. Random assignment of experimental manipulations in a controlled setting allows us to minimize possible confounding and selection effects that otherwise may plague survey-based, field, or naturally occurring data. We can therefore associate treatment effects directly to the experimental manipulations, rather than potentially uncontrolled factors. Accordingly, we conduct our experiment on the platform provided by Amazon Mechanical Turk (MTurk) because it offers the following advantages. First, participant samples on MTurk have been validated to be equally representative to other Internet-based field experiment samples (Buhrmester et al. 2011), and they are consistent with student and community samples

(Goodman et al. 2013). In addition, previous literature has shown that findings from a traditional source (e.g., students, employees) can be replicated using MTurk as the source of participants (Steelman et al. 2014). Furthermore, MTurk enables us to have a high level of control over specifying required participant demographics, such as nationality. As a result, MTurk has been used as a source of participants for many experimental studies in recent years (e.g., Jenkins et al. 2016; Steelman et al. 2014). Next, we describe the measurements we use to evaluate the performance of the proposed password strength meters.

5.1 Measurements

Our first measure evaluates the strength of the password generated, representing the most important purpose of password strength meters. Note that in our study, we measure the increase in password strength after receiving the warning message from the password strength meter rather than just considering final password strength. The choice of our measurement here differentiates our work from other works that only focus on final password strength, as doing so could lead to significantly misleading interpretations of password strength meter effectiveness. For instance, some users initially enter a strong password that is never changed, whereas other users initially enter a weak password but change it in response to password strength meter feedback. These two types of users should not be interpreted as the same in an analysis. If we were to only observe the strength of the final password, it is impossible to distinguish those two aforementioned types of users, and they will be treated similarly (e.g., users with a strong password) in an analysis. Note that password strength follows the power law distribution by nature, so we employ a common practice in the literature (e.g., Vance et al. 2013) by applying the natural logarithm to the value of password strength generated by the Backoff Markov Model.

Second, password strength meters should encourage users to choose strong passwords by nudging them to revise originally entered passwords to those that are stronger. This aspect of password strength meters is crucially important but often overlooked by other password strength meter studies. To highlight this point, consider a password strength meter that is significantly effective in promoting users to think about and enter a strong password. However, users under this meter do not change the originally chosen password but may instead promise themselves they will change the password later. In such a scenario, it cannot be claimed that the password strength meter in question is effective because the password strength has not been revised.

Further, in many contexts—passwords and security behaviors included—individuals become accustomed to engaging in certain behaviors. In the password generation context, individuals are not only accustomed to choosing weak passwords (Zhang et al. 2009; Zviran and Haga 1999), but also reuse those weak passwords across many systems (Ives et al. 2004). Therefore, individuals may require some level of nudging to disengage from the poor password security that they are accustomed to choosing. Prior research has shown that from a policy perspective, "gentle nudges" can be implemented to affect the stickiness of norms of behavior (Kahan 2000). Similarly, nudging has been implemented to affect choice while guiding individuals towards better outcomes (John et al. 2009). As applied to our study, we evaluate whether improved user comprehension of persuasive warning messages will nudge users, motivating them to respond by revising their entered passwords.

Third, in addition to affecting password strength and password revisions, persuasive messages may also encourage users to become more involved in understanding how to create strong passwords to prevent future harm. Prior research has shown that mere awareness of possible security countermeasures is enough to motivate individuals to engage (disengage) in desired (from undesired) behaviors (Anderson and Agarwal 2010; D'Arcy et al. 2009; Spears and Barki 2010). The increased awareness of security also improves the security skills in individuals' practiced behavior (Reddy and Rao 2016). As applied to a password strength meter, efforts to learn more about password security can be measured by tracking user involvement with the meter. That is, in our study, we present a link at the bottom of the password strength meter labeled as "Tips toward strong password". Users that click on the link are educated on how to create stronger passwords, thereby promoting security awareness. Next, we introduce the experimental design and implementation of the password strength meter.

5.2 Experimental Design and Implementation

In this section, we discuss how the study is presented to the participants. Our experiment involves the collection and analysis of passwords, so we are required by the principle of informed consent to disclose to the subjects that they are participating in a study regarding password generation and usage, that passwords generated in this study will be recorded, and that they should not use any of their existing passwords. Although previous literature has shown there is no significant difference in participant behavior in fully disclosed experiments and experiments that employ deception (Bonetti 1998), some concerns could be raised regarding potential biases

that may occur because of such disclosure. For example, it is possible that tasks in the study may not be taken seriously (e.g., randomly picking passwords without an intent to remember). Therefore, we verify that participants remember the password they generated (see details in Section C.6 in the online appendix). In addition, we validate the results through a randomized field experiment to ensure that the results carry over in a more realistic setting (see Section 6).

5.2.1 Hypothetical Scenarios

To ensure that participants experience sufficient realism and purpose, while at the same time maintaining control of the experimental environment, we require that participants create an online account and enter a password using our password strength meter in the process. In this regard, we designed and implemented randomly assigned hypothetical scenarios that required account creation, simultaneously enhancing realism and establishing a purpose for participating in the experiment. Treatments were also randomly assigned, so any systematic bias would be equally likely in each scenario and treatment.

Protection motivation theory (Rogers 1975) and the theory of planned behavior (Ajzen 1985) indicate that users would protect themselves differently based on the potential loss involved. For example, users might react more to the password strength meter when the account they are using stores important information in contrast to how users might react when the account stores unimportant information. In that regard, we derive and implement hypothetical scenarios as follows. At the beginning of the experiment, each participant is presented with one randomly assigned hypothetical scenario from a list of three scenarios: *Forum*, *Restaurant*, and *Bank*. The instructions that participants receive in each scenario are listed in Appendix C. These scenarios were chosen because they each require a different level of sensitivity in the information necessary to setup the online account. The Forum scenario is the simplest of the three and is designed to represent a user that creates an online account to access a bulletin board system (e.g., phpBB). In this scenario, the system only asks for users' email address. The next scenario is Restaurant, and it differentiates from Forum because an online account in the Restaurant scenario requires additional explicit personal information (e.g., name and address). An example of the Restaurant scenario would be setting up an account with a restaurant that provides food delivery services. Lastly, the Bank scenario is the most sensitive of the three, as it requires users to store both financial and

personal information in the online account (e.g., online banking). We also included a clickable button with a caption "Tips toward strong password" at the bottom of the password strength meter in each of the scenarios. If the button is clicked, password generation tips are displayed to the user.[3]

5.2.2 Data Collection

To ensure that the system worked as intended, we first conducted a pilot test and made minor adjustments to the software as necessary. The details presented here are for the full study that followed. For each participant, we recorded an entire history of passwords generated using the password strength meter. All interactions with the system, including sequential events such as clicking the "Tips toward strong password" button, were recorded for use in the analysis.

The MTurk sample recruited for the experiment consisted of 500 participants. At the beginning of the study, participants consented to take part in the experiment. At that time, it was made clear to all participants that passwords would be collected and analyzed in an unencrypted, but anonymized manner. All participants that completed the experiment were compensated $0.75 USD. Although the amount may appear to be very small, it is consistent with payoff guidelines in the literature (e.g., Barger et al. 2011).

Participation was limited to those aged 18 years or older per Amazon Mechanical Turk's terms of use and located in the United States. 74 out of the 500 participants we recruited are from 18 to 22; 202 of them are from 23 to 30; 133 of them are from 31 to 40; 64 of them are from 41 to 50; 25 of them are more than 51 and the remaining two refused to provide their age. Of the 497 participants who were willing to tell us their gender, 283 of them are male and 214 of them are female. The majority of participants have attended some college (no degree), or have a bachelor's degree (i.e., counts of 192 and 205, respectively). For the rest of the participants, 41 are high school graduates; 47 have a master's degree, 8 have a professional degree, and 6 have a doctoral degree. Three participants did not reveal their education level. These participants were randomly assigned to 4 groups and 3 scenarios: 116 are assigned to the control group, 133 are in time treatment, 131 are in rank, and 120 are in probability. As per the scenario assignment, 180 are in bank, 166 are in restaurant, and 150 are in forum. The details of the treatment/scenario assignment, and the post-test survey to ensure the reliability of

[3] See https://support.microsoft.com/en-us/help/4026406/microsoft-account-how-to-create-a-strong-password

participants' data are presented in Appendix C. In addition, details regarding randomization checks of participants in this study are provided in Appendix E.5.

5.3 Results

In this section, we present the results of our controlled laboratory experiment. Here, we utilize a regression framework with the following specification as our primary method of analysis:

$$DV = \alpha_0 + \alpha_1 Age + \alpha_2 Gender + \alpha_3 Education + \beta_1 Time + \beta_2 Rank + \beta_3 Probability + \gamma_1 Restaurant + \gamma_2 Forum + \varepsilon,$$

where DV represents three dependent variables described in Section 6.1. α_1 to α_3 capture the effect of the control variables, β_1 to β_3 capture the effect of the persuasive messages, using the traditional password strength meter (i.e., no persuasive message) as a baseline, γ_1 and γ_2 capture the effect of hypothetical scenarios, using the bank scenario as a baseline. The treatment variables (i.e., $Time$, $Rank$, and $Probability$) are binary variables that take the value one for participants who observe such treatments and zero otherwise. For example, $Time = 1$ for participants who observe the password strength meter with fear appeals message. Meanwhile, our demographic survey question regarding participants' age is in a "bin" format so we use the middle of each bin as the age variable in the analysis, and gender is a dummy variable where 0 represents male participants and 1 represents female participants. Lastly, education is the number of years required to complete the education level that the participant chose (e.g., 12 for a high school diploma, 16 for a bachelor's degree, etc.) Note that the first dependent variable (*diff_strength*), representing the increase in the natural logarithm of the password strength before/after observing the warning message from the password strength meter, is a continuous variable. So, we report the results using linear regression analysis. The second dependent variable (*num_reset*) represents the number of occasions that participants change their passwords immediately after observing the warning message from the password strength meter. Because it is a count variable, we report the results using a Poisson regression analysis. Lastly, the third dependent variable (*learn_more*) is a binary variable that takes a value of one for participants who visit the link to learn more about password security and zero otherwise. Therefore, we report the results using a logistic regression analysis.

Table 3: Results from the regression analyses

	diff_strength	*num_reset*	*learn_more*
Treatment			
time	0.031 (0.028)	0.288 (0.357)	-0.263 (0.355)
rank	**0.089** (0.028)**	**0.993** (0.322)**	-0.510 (0.370)
probability	0.046 (0.029)	0.107 (0.339)	-0.396 (0.373)
Scenario			
restaurant	0.047 (0.024)	0.370 (0.251)	0.005 (0.306)
forum	0.063* (0.024)	0.389 (0.255)	-0.304 (0.303)
Control			
age	-0.001 (0.010)	-0.079 (0.100)	0.017 (0.128)
gender	0.011 (0.020)	-0.119 (0.209)	-0.050 (0.267)
education	0.002 (0.011)	-0.147 (0.119)	0.045 (0.142)
Constant	-0.039 (0.048)	-1.754*** (0.524)	-1.628** (0.618)
Observations	497	497	497
R-squared	0.034	0.033	0.008

Notes. Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$

Table 3 reports the results from our regression analyses. First, we find that none of the control variables have a statistically significant influence at $p < 0.05$ on any of the dependent variables. For the difference in password strength, our analysis reveals that although the average increases in password strength of all 3 types of persuasive messages are higher than the control, only the difference of *rank* is statistically significant. The next largest increase in password strength is observed in the *probability* treatment, followed by the *time* treatment. Interestingly, contrary to conventional wisdom, the password strength meters work better in the *forum* scenario compared to the *bank* scenario in terms of inducing users to create a stronger password. As for the effect of our proposed password strength meter on the number of occasions where users change their password after observing a warning message provided by the password strength meter, our results are consistent with the results on the increase in password strength. We observe that all 3 types of persuasive message yield a higher number of occasions where users change their password in comparison to the control group where no persuasive message is provided. However, only the difference of *rank* is statistically significant at $p < 0.01$. In addition, the scenario does not have any significant effect here. Lastly, the results show that there is no significant difference between the three treatment groups and the control group in terms of the number of times where users click the password generation tips. Note that the analyses here only contain 497 observations

because we lost 3 participants who refused to answer the demographic questions. However, the results are qualitatively similar when we include those participants and exclude control variables from the analyses.

Next, we expand on the interesting result where the average increase in password strength is lowest in the bank scenario. Recall that user accounts in the bank scenario store both personal information and financial information. Therefore, one would intuitively expect that users in the bank scenario would be more concerned about the security of their account and thus would be more responsive to the proposed password strength meters in comparison with the restaurant scenario (only personal information stored), and the forum scenario (no personal/financial information stored). However, it turns out that our proposed password strength meter works significantly better in terms of inducing users to select stronger passwords for the forum scenario, which has the lowest level of security concerns. It is worth noting that we also observe a similar pattern for the difference in the number of occasions that users change their password after observing the warning message generated by password strength meters (i.e., users under bank scenario change their passwords less often in comparison to users in the other two scenarios). However, the difference in that metric is not statistically significant at $p < 0.05$. A possible explanation for this counter-intuitive result is that users tend to believe that financial-related websites usually offer other techniques to protect their account (e.g., sending a confirmation code to their cell phones, fraud detection techniques, etc.), so they do not tend to set a strong password in such a scenario. Alternatively, one may argue that users in the bank scenario may initially generate a more secure password in comparison to users in other scenarios. As a result, there is less room for improvement after observing the warning message displayed by the password strength meter. Therefore, to control for such an effect, we add another control variable, *initial_strength_label*, which is the label of the first password generated observed by each user, to the model and rerun the regression analyses on all of our dependent variables.

The results, reported in Table 4, show that the initial password strength indeed plays an important role in determining the increase in password strength and the number of occasions that users change their password after observing the warning message generated by password strength meters. However, even when the initial password strength is included in the model, the coefficients of both the average increase in password strength

and the number of occasions that users change their password after observing the warning message generated by password strength meters across scenarios are still positive and statistically significant.

Table 4: Results from the regression analyses (controlling for the initial strength label)

	diff_strength	*num_reset*	*learn_more*
Treatment			
time	0.030 (0.027)	0.286 (0.358)	-0.507 (0.371)
rank	**0.089** (0.027)**	**0.980** (0.322)**	-0.260 (0.355)
probability	0.043 (0.028)	0.107 (0.339)	-0.389 (0.373)
Scenario			
restaurant	0.039 (0.023)	0.301 (0.252)	0.019 (0.307)
forum	0.045 (0.024)	0.289 (0.259)	-0.271 (0.336)
Control			
age	0.004 (0.009)	-0.054 (0.100)	0.010 (0.129)
gender	0.009 (0.019)	-0.109 (0.210)	-0.046 (0.267)
education	0.002 (0.010)	-0.156 (0.121)	0.044 (0.142)
initial_strength_label	-0.086*** (0.014)	-0.400** (0.125)	0.155 (0.209)
Constant	0.183** (0.059)	-0.746 (0.614)	-2.030* (0.823)
Observations	497	497	497
R-squared	0.103	0.050	0.010

Notes. Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$

In addition to the main models presented above, we also conduct several additional analyses as follows. First, one may argue that some participants may not generate a strong password because they do not know how to do so. To control for this potential confounding factor, we add the variable *learn_more* as another control variable to the model that examines *diff_strength* and *num_reset*. The results, presented in Appendix E.1, are qualitatively similar to our main results. Second, we demonstrate that the variable *num_reset* is a mediator of the effect of proposed password strength meters on *diff_strength*. The complete results of the mediation analysis are presented in Appendix E.2. Third, we consider the moderation role of *initial_strength_label* on the treatment effects, where the results are reported in Appendix E.3. We find that *initial_strength_label* indeed moderates the effect of proposed password strength meters on *diff_strength* but the moderation effects on *num_reset* and *learn_more* is not statistically significant. Fourth, we consider an alternative password strength calculation using password entropy. The results, which are qualitatively similar to our main results, are reported in Appendix E.4.

In summary, the randomized laboratory experiment demonstrates that our ELM augmented password strength meter is significantly more effective than the traditional one in terms of increasing the strength of the

password and the number of occasions the user changes the password immediately after observing the meter for the rank treatment (i.e., peer comparison persuasive message). However, there could be some concerns regarding the external validity of the results, primarily because of the informed consent part of the experiment where users are aware that they are being observed in an experiment regarding passwords. Further, the stakes for participants are lower in lab studies as compared to field studies, which could lead to conservative results due to less cognitive effort for participants (e.g., Levitt and List 2007). In a lab experiment, the only thing at stake is earning a payoff, whereas in a field experiment a participant's account could be compromised. That aspect may lead to higher cognitive effort in comparison to a lab experiment and thus the central route of the ELM could be more likely to be followed. To alleviate such concerns, we next conduct a randomized field experiment to further validate the results in a realistic setting.

6. Study 3: Validating the Design in a Real-World Setting

In the third study, we seek to establish the validity of the results we observe in the laboratory experiment. Doing so would constitute "going the last research mile" in establishing proof-of-use of our design (Nunamaker et al. 2015). In that regard, we collaborated with a small website in Asia to conduct a field experiment. The partner website is an online forum that provides a platform for users to discuss local deals and promotions. More than 95% of the visitors are from the country in which the website operates. Consistent with our definition of the forum scenario in the laboratory experiment, the platform only requires the user to provide a valid email address. No other personal information (e.g., name, gender, address, etc.) is required in the registration process. The experimental design is similar to the laboratory experiment described in Section 5. That is, users who visit the partner website and register for a new account during the experimental period are randomized into one of the four treatments (i.e., control, time, rank, probability).[4] Obviously, the corresponding scenario is only *forum* and we do not impose any other scenarios in the field experiment. The rest of the experimental settings (e.g., the persuasive message, the way we calculate the password strength, etc.) are identical to the laboratory

[4] Users are assigned to the treatment condition based on their session ID. Particularly, the platform performed the *div* operation on the last bit of the session ID with the value 4. As such, the result (i.e., remaining quotient) is either 0, 1, 2, or 3, each of which corresponds to a treatment condition (0=control, 1=time, 2=rank, 3=probability).

counterpart. Screenshots of each type of persuasive message in the field experiment are provided in Appendix D. Also, we are only interested in two password-generating behaviors that show signs of improvement through our proposed password strength meter. First, the increase in password strength; and second, the number of occasions that the users change their password immediately after observing the meter.

One important detail to note regarding the field experiment is that our partner website operates in a country where law prohibits the collection of personal and sensitive information (including passwords) unless the users are sufficiently notified. This restriction creates a dilemma for us because notifying the users regarding our research activity would defeat the purpose of running a field experiment to validate the proof-of-use of our proposed design in the first place. Therefore, to maximize the realism of this field experiment, we utilize the following approaches, which allow us to execute the experiment without notifying the users while still complying with the law. First, instead of collecting the password as in the laboratory experiment, we compute the strength of every password generated at run-time and only collect the password strength, which is neither personally identifiable nor sensitive. Second, we worked with the website to develop a separate system based on the concept of the Chinese wall model in information security (Gordon 2015). Particularly, once a user generates a password on the website, the website's system would calculate the strength of that password and send it to the researcher system without storing it locally. This way, the website does not even observe the strength of the password generated while the researcher does not observe anything else apart from the password strength (and the number of occasions that the user changes her password during the account creation). The downside of our approach is that we can only obtain a system-generated ID (which is different from the user ID that the website is using) and the corresponding password strength for every time the user enters the password on the account creation page (i.e., we are unable to obtain any other user information or observe any user activities on the website). Nevertheless, the website provides us two application program interfaces (API) that allow us to verify the validity of the passwords generated in this study while maintaining legal compliance. The first API enables us to verify whether the user logs in to the website within the first 24 hours after account creation. This would allow us to exclude users who register for an account without using it because their password-generating behavior might not reflect natural password-generating behavior. The second API enables

us to verify whether the user uses the "forgot password" function within 7 days after account creation. The use of this function may indicate that the users did not pay attention when they generated their passwords and hence, they should be removed from the study.

The experiment was implemented on the partner website for 30 days. In total, 310 users created a new account on our partner website during this period. All the users logged in to the website at least once within 24 hours after account creation. However, two of the users were unable to recall the password within 7 days after joining the website so we removed them from the analysis. As a result, we have 308 users in total: 76 users for the control group, 79 users for the time treatment, 78 users for the rank treatment, and 75 users for the probability treatment. As previously mentioned, unlike the analysis for the laboratory experiment, we are unable to include any control variables for this analysis due to limitations of the field setting (although note that the effect of the control variables is minimal in the laboratory experiment). Therefore, we leverage an ANOVA model to examine whether there is any difference on our variables of interest between our proposed password strength meter and the traditional password strength meter. Similar to the laboratory experiment, we apply a natural logarithm to the password strength. The increase in password strength, which is a continuous variable, is calculated by differentiating the strength of the password before vs. the strength after observing the password strength meter. Meanwhile, the number of occasions where users change their password after observing a warning message provided by the password strength meter is a count measure. The treatment variables are a binary variable that takes the value one if the participant observes the password strength meter that utilizes the persuasive message of that treatment and zero otherwise.

The results from Table 5 demonstrate that there is at least one type of persuasive message that is significantly more effective than the traditional password strength meter in terms of inducing users to generate stronger passwords and persuading users to change their password after observing the message. We also consider the moderating role of *initial_password_strength* and report the results, which are qualitatively similar to those from the laboratory experiment, in Appendix E.3. In addition, we also consider a model that controls for *initial_password_strength* in Appendix E.6.

Table 5: Results from the ANOVA analyses

		Sum of squares	**d.f.**	**Mean square**	***F*-ratio**	**Significance**
Increase in password strength	Between groups	1.533	3	0.511	4.11	0.007
	Within groups	37.820	304	0.124		
	Total	39.353	307	0.128		
Revisions to Entered Passwords	Between groups	7.361	3	2.454	4.83	0.003
	Within groups	155.467	304	0.508		
	Total	163.828	307	0.527		

Next, we perform post-hoc pairwise comparison tests to examine the effectiveness of each treatment in comparison to the control treatment. Here, we adjust our *p*-values following the Bonferroni correction to address the issue of multiple comparisons to ensure that our results remain conservative.

Table 6: Post hoc pairwise comparisons tests (*p*-value adjusted by Bonferroni correction)

Comparison	**Increase in password strength**		**Revisions to Entered Passwords**	
	Mean difference	**Significance**	**Mean difference**	**Significance**
Time vs. Control	0.1458	0.065	0.2935	0.067
Rank vs. Control	0.1861	0.007	0.3928	0.004
Probability vs. Control	0.1408	0.077	0.0993	1.000
Probability vs. Time	-0.0050	1.000	-0.1942	0.546
Probability vs. Rank	-0.0452	1.000	-0.2935	0.067
Rank vs. Time	0.0402	1.000	0.0993	1.000

Table 6 reports the results of the pairwise comparisons. Overall, the results are consistent with what we observe in the laboratory experiment. The *rank* treatment is significantly effective in terms of inducing users to both generate strong passwords and encouraging them to do so more often, compared to the traditional password strength meter. Meanwhile, we find that the effectiveness of other two treatments (i.e., *time* and *probability*) become statistically significant in the field experiment. Nevertheless, it is important to note that the significance level of *time* and *probability* is only marginal (i.e., $p < 0.10$) and that the field experiment does not have control variables so their effects must be interpreted cautiously. Note also that the *probability* treatment does not perform better than the traditional password strength meter in terms of encouraging users to change their password. Yet, when the users do change their password, the *probability* treatment is effective in inducing them to choose a stronger one.

7. Discussions and Conclusions

Persuading users to use strong passwords has consistently been an important task for firm managers and researchers. Our study aids such effort by adopting the multimethod approach to enhance the effectiveness of password strength meters. We first draw upon the psychology literature to identify an underpinning theory that can help guide the theoretically augmented design of password strength meters. We find that the Elaboration Likelihood Model of Persuasion (ELM) fits well with our framework as it explains potential ineffectiveness of traditional password strength meters and also offers a guideline that we use to develop our conceptual design. Particularly, our augmented password strength meter design incorporates persuasive messages that are likely to be processed through the central route according to ELM. As a result, changes in behavior are likely to be more salient, compared with the change in the behavior through traditional password strength meters which tends to persuade users through the peripheral route processing. In our design, we rely on three types of persuasive message: fear appeals, peer comparisons, and common bond.

Research Summary

We conducted three experiments that utilize different research methodologies (survey; controlled laboratory experiment; field experiment) to evaluate our proposed design and confirm that the performance of existing password strength meters can be enhanced by incorporating meaningful persuasive messages such as the rank of the entered password in comparison to other common passwords. Overall, the results of our studies are complementary to each other, and provide both theoretical and managerial contributions to the literature. The results from Study 1 show that the effect of persuasive messages on user behavior appear to be consistent with the meaning of the message, providing evidence that users carefully and thoughtfully comprehend the implications of strong passwords (i.e., that the messages are likely to be processed through the central route according to ELM). The results from Studies 2 and 3 are consistent with each other as regards to the order of the treatments on their effects on password-generating behavior. The rank treatment is the best performing for both studies, while the other two treatments have a marginally significant effect on behavior in Study 3 and statistically insignificant in Study 2. In this regard, we realize the benefits of going the last research mile by implementing a proposed password strength meter in the field, as it provides the strongest external validity in

experimental design. Further, when we compare the contributions of our work to the literature, we address an open issue of identifying a root cause of password ineffectiveness and the mechanisms that can improve password strength without creating new algorithms, which we expand upon next.

Theoretical Contributions

Our research makes two important theoretical contributions. First, our paper contributes to the IS literature on password strength meters by providing a theoretical explanation of how password strength meters influence user behaviors. Drawing on ELM, we theorize how different stimuli might be processed by users through the peripheral or central route, and consequently, how they might differently affect user password behaviors. Our theoretical development helps add explainability to the design of password strength meters, as aspect that prior works note to be lacking in the literature (Carnavalet and Mannan 2014; Carnavalet and Mannan 2015). Our paper can also serve as the theoretical foundation for future work on password strength meters. Our adaptation of ELM to the digital context can help to better understand what stimuli might be processed through peripheral and central routes and how this affects the likelihood of improving password-generating behavior. Furthermore, our paper provides evidence of the importance of conducting research that focuses on the use of persuasive messages. Our results indicate that without changing the password strength meter algorithm, users' password-generating behaviors can be significantly improved by simply adding theory-driven persuasive messages to induce users to carefully and comprehensively process the information supplied by the password strength meter. Hence, making theoretically-based minor changes to the password strength meter could potentially lead to stronger passwords in general.

Second, our paper contributes to the ELM literature as well. By adapting ELM and applying it in a digital context, our study suggests that the effectiveness of stimuli might be context dependent, which has not been previously acknowledged. Engagement in the digital world is highly influenced by social and community-based factors (e.g., Aggarwal and Singh 2013; Chen et al. 2011) where individuals share and compare information and advice (Huang et al. 2019), and where decisions are influenced by social comparisons such as reviewers' assessments and opinions about products and services or popularity of products (Dewan et al. 2017; Oh et al. 2016). Our study demonstrates that while our three messages parallel those of ELM for central route processing,

in the digital context, peer comparison outperforms the other type of stimuli in terms of inducing users to improve their password-generating behavior. The result suggests that in a digital context, where social comparison is important, such a construct is best because it embeds a comparison to other users that are in a similar situation, which may create context and bounds for the user's attention and cognitive ability. Conversely, it is possible that the other two treatments provide stimuli that are not aligned with the contextual expectations of individuals and therefore require additional cognitive ability to process and/or unbounded contexts, which may cause the messages to be less successful at persuading change for the users. Our findings have important implications for ELM research in that it can lead to revisiting premises about the importance of different types of stimuli for nudging behaviors.

Managerial Implications

The findings presented in this paper also provide three relevant and important managerial implications. First, the increase in password strength after users observe the warning message from a proposed password strength meter is significant both statistically and economically. Given that user accounts are a common channel that tend to be exploited, such an increase in password strength provides a significant additional safeguard for firms from a common cyber-attack such as brute forcing which takes advantage of weak passwords. To put it in perspective, the 18% increase in password strength that we observed in study 3 (under the *rank* treatment) means that a hacker would need approximately 5 more weeks to successfully brute force an average account in our dataset, based on the estimated time to crack calculation in Vance et al. (2013). Although it is difficult to quantify monetary values from such an improved behavior, an overall increase in end-point security is widely appreciated by firm managers because it can help prevent the disruption in firm operations and complicated legal and trust issues that ensue if unauthorized access occurs. This practical implication of our work is especially important given that the majority of popular websites still rely on traditional password strength meters to improve users' password-generating behavior. Our implementation nudges users to create strong passwords without requiring complicated password complexity to encourage users to create strong passwords that are easier to remember. Second, unlike many advanced technologies that can alleviate an issue of weak passwords (e.g., authentication manager software, password complexity policies, etc.), the proposed methodology to

embed persuasive messages into an existing password strength meter requires minimal financial and technical investment. The proposed method works with any algorithm used by existing password strength meters. Also, the computational task is performed on the client-side. Hence, no additional computational resources are required on the firm side. Yet, the results demonstrate that proposed password strength meters are significantly more effective than traditional meters in multiple dimensions. Third, the results from our laboratory experiment demonstrates that embedding password rank in the password strength meter is not only more effective than the traditional password strength meter, but also more effective at promoting user secure behavior in a not-so-secure environment where users initially have fewer concerns regarding information security. These findings are important because many websites do not require users to store their personal and financial information. However, those websites are often a target of hackers, which can lead to cascading attacks on related accounts on other websites. Also, hackers can use the data from those websites to craft a personalized attack (e.g., spear phishing) in an attempt to compromise accounts with more important data. In that regard, our proposed password strength meter can aid firm managers to help end-users create stronger passwords in less sensitive contexts.

Limitations and Future Research

Our research is not without limitations. Our first experiment is exploratory and is only concerned with data reliability and observed differences between treatments when using existing theoretical constructs. Next, Amazon Mechanical Turk participants are known to be tech-savvy, which might introduce bias to the password strength observed in our second experiment. Nevertheless, it is important to note that the potential bias in study 2 does not affect the ability of our treatment manipulations to nudge behavior due to the controls implemented there. Relatedly, we include the "Tips toward strong password" link as the dependent variable but did not directly study its detailed implications. For instance, what happens after users click that link? Do they really read the provided information? Can they remember that information and use it later? Our design intentionally did not capture such information or if participants wrote down or reused their password, so future research could employ additional surveys or user tracking to gauge understanding before and after the study and how these changes may affect password-generating behavior. Moreover, the three types of persuasive

messages we proposed are based on prior literature and are not exhaustive. Future research could examine other types of persuasive messages that may improve password-generating behavior. In addition, the password strength calculation in this study is primarily calculated by using the RockYou dataset as a training dataset. If would be interesting for future research that has access to a more recent plaintext dataset to verify whether the change in password-generating behavior affects the password strength calculation. Lastly, our study does not compare the effectiveness of the proposed password strength meters with other non-traditional password strength meters. It would be interesting for future meta research to collect and compare the effectiveness of different approaches to enhance the effectiveness of password strength meters under similar settings.

Acknowledgements

The authors would like to thank the senior editor, the associate editor, and two anonymous reviewers for their helpful feedback in the review process. The authors also thank Hyunji So for her helpful suggestion. In addition, we are grateful for participants at HICSS 2017, WITS 2017, ISOM Research Seminar at George Mason University, and the IS Research Seminar at Chinese University of Hong Kong for their constructive comments and discussions. Lastly, we gratefully acknowledge financial support from the Social Science and Humanities Research Council of Canada (SSHRC) grant number 435-2018-0605.

References

Aggarwal, R., and Singh, H. 2013. "Differential Influence of Blogs across Different Stages of Decision Making: The Case of Venture Capitalists," *Mis Quarterly*), pp. 1093-1112.

Ahn, H. Y., Park, J. S., and Haley, E. 2014. "Consumers' Optimism Bias and Responses to Risk Disclosures in Direct-to-Consumer (Dtc) Prescription Drug Advertising: The Moderating Role of Subjective Health Literacy," *Journal of Consumer Affairs* (48:1), pp. 175-194.

Ajzen, I. 1985. "From Intentions to Actions: A Theory of Planned Behavior," in *Action Control*. Springer, pp. 11-39.

Alegre Ansuategui, F. J., and Moliner Miravet, L. 2017. "Emotional and Cognitive Effects of Peer Tutoring among Secondary School Mathematics Students," *International Journal of mathematical education in science and technology* (48:8), pp. 1185-1205.

Anderson, C. L., and Agarwal, R. 2010. "Practicing Safe Computing: A Multimedia Empirical Examination of Home Computer User Security Behavioral Intentions," *MIS quarterly* (34:3), pp. 613-643.

Angst, C. M., and Agarwal, R. 2009. "Adoption of Electronic Health Records in the Presence of Privacy Concerns: The Elaboration Likelihood Model and Individual Persuasion," *MIS quarterly* (33:2), pp. 339-370.

Bachmann, M. 2014. "Passwords Are Dead: Alternative Authentication Methods," *2014 IEEE Joint Intelligence and Security Informatics Conference*: IEEE, pp. 322-322.

Barger, P., Behrend, T., Sharek, D., and Sinar, E. 2011. "Io and the Crowd: Frequently Asked Questions About Using Mechanical Turk for Research," *The Industrial-Organizational Psychologist* (49:2), pp. 11-17.

Bhattacherjee, A., and Sanford, C. 2006. "Influence Processes for Information Technology Acceptance: An Elaboration Likelihood Model," *MIS quarterly*), pp. 805-825.

Bolton, R. 2014. "The Persuasive Pressure of Peer Rankings," in: *Harvard Business Review*.

Bonetti, S. 1998. "Experimental Economics and Deception," *Journal of Economic Psychology* (19:3), pp. 377-395.

Boss, S., Galletta, D., Lowry, P. B., Moody, G. D., and Polak, P. 2015. "What Do Systems Users Have to Fear? Using Fear Appeals to Engender Threats and Fear That Motivate Protective Security Behaviors," *MIS Quarterly (MISQ)* (39:4), pp. 837-864.
Bowes, R. 2015. "Passwords-Skullsecurity." from http://wiki.skullsecurity.org
Buhrmester, M., Kwang, T., and Gosling, S. D. 2011. "Amazon's Mechanical Turk: A New Source of Inexpensive, yet High-Quality Data?," *Perspectives on psychological science* (6:1), pp. 3-5.
Carnavalet, X. D. C. D., and Mannan, M. 2014. "From Very Weak to Very Strong: Analyzing Password-Strength Meters," *Network and Distributed System Security Symposium (NDSS 2014)*: Internet Society.
Carnavalet, X. D. C. D., and Mannan, M. 2015. "A Large-Scale Evaluation of High-Impact Password Strength Meters," *ACM Transactions on Information and System Security (TISSEC)* (18:1), pp. 1-32.
Castelluccia, C., Dürmuth, M., and Perito, D. 2012. "Adaptive Password-Strength Meters from Markov Models," *NDSS*.
Chen, Y., Wang, Q., and Xie, J. 2011. "Online Social Interactions: A Natural Experiment on Word of Mouth Versus Observational Learning," *Journal of marketing research* (48:2), pp. 238-254.
Cho, H., Lee, J.-S., and Chung, S. 2010. "Optimistic Bias About Online Privacy Risks: Testing the Moderating Effects of Perceived Controllability and Prior Experience," *Computers in Human Behavior* (26:5), pp. 987-995.
Chung, N., and Han, H. 2017. "The Relationship among Tourists' Persuasion, Attachment and Behavioral Changes in Social Media," *Technological Forecasting and Social Change* (123), pp. 370-380.
Cimpanu, C. 2018. "State Department Shamed for Poor Adoption of Multi-Factor Authentication." Retrieved May 17, 2019, from https://www.zdnet.com/article/state-department-shamed-for-poor-adoption-of-multi-factor-authentication
Cyr, D., Head, M., Lim, E., and Stibe, A. 2018. "Using the Elaboration Likelihood Model to Examine Online Persuasion through Website Design," *Information & Management* (55:7), pp. 807-821.
D'Arcy, J., Hovav, A., and Galletta, D. 2009. "User Awareness of Security Countermeasures and Its Impact on Information Systems Misuse: A Deterrence Approach," *Information systems research* (20:1), pp. 79-98.
Dewan, S., Ho, Y.-J., and Ramaprasad, J. 2017. "Popularity or Proximity: Characterizing the Nature of Social Influence in an Online Music Community," *Information Systems Research* (28:1), pp. 117-136.
Dinoff, B. L., and Kowalski, R. M. 1999. "Reducing Aids Risk Behavior: The Combined Efficacy of Protection Motivation Theory and the Elaboration Likelihood Model," *Journal of Social and Clinical Psychology* (18:2), pp. 223-239.
Dunbar, N. E., Connelly, S., Jensen, M. L., Adame, B. J., Rozzell, B., Griffith, J. A., and Dan O'Hair, H. 2014. "Fear Appeals, Message Processing Cues, and Credibility in the Websites of Violent, Ideological, and Nonideological Groups," *Journal of Computer-Mediated Communication* (19:4), pp. 871-889.
Egelman, S., Sotirakopoulos, A., Muslukhov, I., Beznosov, K., and Herley, C. 2013. "Does My Password Go up to Eleven? The Impact of Password Meters on Password Selection," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2379-2388.
Ernst, J. M., and Heesacker, M. 1993. "Application of the Elaboration Likelihood Model of Attitude Change to Assertion Training," *Journal of counseling psychology* (40:1), p. 37.
Festinger, L. 1954. "A Theory of Social Comparison Processes," *Human Relations* (7:2), pp. 117-140.
Furnell, S. 2011. "Assessing Password Guidance and Enforcement on Leading Websites," *Computer Fraud & Security* (2011:12), pp. 10-18.
Furnell, S., Khern-am-nuai, W., Esmael, R., Yang, W., and Li, N. 2018. "Enhancing Security Behaviour by Supporting the User," *Computers & Security* (75), pp. 1-9.
Gefen, D., and Straub, D. 2005. "A Practical Guide to Factorial Validity Using Pls-Graph: Tutorial and Annotated Example," *Communications of the Association for Information systems* (16:1), p. 5.
Goecks, J., and Mynatt, E. D. 2005. "Supporting Privacy Management Via Community Experience and Expertise," in *Communities and Technologies 2005*. Springer, pp. 397-417.
Golla, M., and Dürmuth, M. 2018. "On the Accuracy of Password Strength Meters," *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1567-1582.
Golla, M., Hahn, B., zu Selhausen, K. M., Hosseini, H., and Dürmuth, M. 2018. "Bars, Badges, and High Scores: On the Impact of Password Strength Visualizations," *Who Are You?! Adventures in Authentication (WAY)*, Baltimore, MD, USA.
Goodman, J. K., Cryder, C. E., and Cheema, A. 2013. "Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples," *Journal of Behavioral Decision Making* (26:3), pp. 213-224.
Gordon, A. 2015. *Official (Isc) 2 Guide to the Cissp Cbk*. CRC Press.
Grassi, P. A., Garcia, M. E., and Fenton, J. L. 2017. "Nist Special Publication 800-63-3 Digital Identity Guidelines," *National Institute of Standards and Technology, Los Altos, CA*).
Gregor, S. 2006. "The Nature of Theory in Information Systems," *MIS quarterly*), pp. 611-642.
Gregor, S., and Hevner, A. R. 2013. "Positioning and Presenting Design Science Research for Maximum Impact," *MIS quarterly*), pp. 337-355.

Herath, T., and Rao, H. R. 2009. "Protection Motivation and Deterrence: A Framework for Security Policy Compliance in Organisations," *European Journal of Information Systems* (18:2), pp. 106-125.
Houshmand, S., and Aggarwal, S. 2012. "Building Better Passwords Using Probabilistic Techniques," *Proceedings of the 28th Annual Computer Security Applications Conference*, pp. 109-118.
Huang, K.-Y., Chengalur-Smith, I., and Pinsonneault, A. 2019. "Sharing Is Caring: Social Support Provision and Companionship Activities in Healthcare Virtual Support Communities," *MIS quarterly* (43:2), pp. 395-424.
Ives, B., Walsh, K. R., and Schneider, H. 2004. "The Domino Effect of Password Reuse," *Communications of the ACM* (47:4), pp. 75-78.
Jenkins, J. L., Anderson, B. B., Vance, A., Kirwan, C. B., and Eargle, D. 2016. "More Harm Than Good? How Messages That Interrupt Can Make Us Vulnerable," *Information Systems Research* (27:4), pp. 880-896.
Jenkins, J. L., Grimes, M., Proudfoot, J. G., and Lowry, P. B. 2014. "Improving Password Cybersecurity through Inexpensive and Minimally Invasive Means: Detecting and Deterring Password Reuse through Keystroke-Dynamics Monitoring and Just-in-Time Fear Appeals," *Information Technology for Development* (20:2), pp. 196-213.
John, P., Smith, G., and Stoker, G. 2009. "Nudge Nudge, Think Think: Two Strategies for Changing Civic Behaviour," *The Political Quarterly* (80:3), pp. 361-370.
Johnston, A. C., and Warkentin, M. 2010. "Fear Appeals and Information Security Behaviors: An Empirical Study," *MIS quarterly* (34:3), pp. 549-566.
Kahan, D. 2000. "Gentle Nudges Vs Hard Shoves: Solving the Sticky Norms Problem'(2000)," *University of Chicago Law Review* (67:3), pp. 607-645.
Kariryaa, A., and Schöning, J. 2020. "Moiprivacy: Design and Evaluation of a Personal Password Meter," *19th International Conference on Mobile and Ubiquitous Multimedia*, pp. 201-211.
Karjalainen, M., Sarker, S., and Siponen, M. 2019. "Toward a Theory of Information Systems Security Behaviors of Organizational Employees: A Dialectical Process Perspective," *Information Systems Research* (30:2), pp. 687-704.
Kelley, P. G., Komanduri, S., Mazurek, M. L., Shay, R., Vidas, T., Bauer, L., Christin, N., Cranor, L. F., and Lopez, J. 2012. "Guess Again (and Again and Again): Measuring Password Strength by Simulating Password-Cracking Algorithms," *2012 IEEE symposium on security and privacy*: IEEE, pp. 523-537.
Levitt, S., and List, J. 2007. "Viewpoint: On the Generalizability of Lab Behaviour to the Field," *Canadian Journal of Economics* (40:2), pp. 347-370.
Ma, J., Yang, W., Luo, M., and Li, N. 2014. "A Study of Probabilistic Password Models," *2014 IEEE Symposium on Security and Privacy*: IEEE, pp. 689-704.
Maida, K. 2016. "What the New Nist Guidelines Mean for Authentication." Retrieved August 15, 2017, from https://auth0.com/blog/what-the-new-nist-guidelines-mean-for-authentication
Mirante, D., and Cappos, J. 2013. "Understanding Password Database Compromises," *Dept. of Computer Science and Engineering Polytechnic Inst. of NYU, Tech. Rep. TR-CSE-2013-02*).
Munoz, Y., Chebat, J.-C., and Suissa, J. A. 2010. "Using Fear Appeals in Warning Labels to Promote Responsible Gambling among Vlt Players: The Key Role of Depth of Information Processing," *Journal of Gambling Studies* (26:4), pp. 593-609.
Nunamaker, J. F., Briggs, R. O., Derrick, D. C., and Schwabe, G. 2015. "The Last Research Mile: Achieving Both Rigor and Relevance in Information Systems Research," *Journal of management information systems* (32:3), pp. 10-47.
Oh, H., Animesh, A., and Pinsonneault, A. 2016. "Free Versus for-a-Fee: The Impact of a Paywall on the Pattern and Effectiveness of Word-of-Mouth Via Social Media," *MIS Quarterly* (40:1), pp. 31-56.
Ophoff, J., and Janowski, M. 2015. "Gamification as a Driver of Individual Information Security Behavior," *Proceedings of the 2015 Dewald Roode Workshop on Information Systems Security Research*.
Petty, R. E., Baker, S. M., and Gleicher, F. 1991. "Attitudes and Drug Abuse Prevention: Implications of the Elaboration Likelihood Model of Persuasion," *Persuasive communication and drug abuse prevention*), pp. 71-90.
Petty, R. E., and Cacioppo, J. T. 1986. "The Elaboration Likelihood Model of Persuasion," in *Communication and Persuasion*. Springer, pp. 1-24.
Petty, R. E., Cacioppo, J. T., and Schumann, D. 1983. "Central and Peripheral Routes to Advertising Effectiveness: The Moderating Role of Involvement," *Journal of consumer research* (10:2), pp. 135-146.
Pinkas, B., and Sander, T. 2002. "Securing Passwords against Dictionary Attacks," *Proceedings of the 9th ACM conference on Computer and communications security*, pp. 161-170.
Postmes, T., and Spears, R. 2000. "Refining the Cognitive Redefinition of the Group: Deindividuation Effects in Common Bond Vs. Common Identity Groups," *SIDE effects centre stage: Recent developments in studies of de-individuation in groups.*, pp. 63-78.
Reddy, D., and Rao, V. 2016. "Cybersecurity Skills: The Moderating Role in the Relationship between Cybersecurity Awareness and Compliance," *American Conference on Information Systems*.
Reinhard, M.-A., and Sporer, S. L. 2010. "Content Versus Source Cue Information as a Basis for Credibility Judgments," *Social Psychology*).

Ringle, C. M., Wende, S., and Becker, J. 2015. "Smartpls 3. Bönningstedt: Smartpls, 2015."
Rogers, R. W. 1975. "A Protection Motivation Theory of Fear Appeals and Attitude Change1," *The journal of psychology* (91:1), pp. 93-114.
Rubinstein, A. 1988. "Similarity and Decision-Making under Risk (Is There a Utility Theory Resolution to the Allais Paradox?)," *Journal of economic theory* (46:1), pp. 145-153.
Ruiter, R. A., Abraham, C., and Kok, G. 2001. "Scary Warnings and Rational Precautions: A Review of the Psychology of Fear Appeals," *Psychology and health* (16:6), pp. 613-630.
Shay, R., Bauer, L., Christin, N., Cranor, L. F., Forget, A., Komanduri, S., Mazurek, M. L., Melicher, W., Segreti, S. M., and Ur, B. 2015. "A Spoonful of Sugar? The Impact of Guidance and Feedback on Password-Creation Behavior," *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 2903-2912.
Slain, M. 2016. "Announcing Our Worst Passwords of 2016." Retrieved August 15, 2017, from https://www.teamsid.com/worst-passwords2016/
Slovic, P., Monahan, J., and MacGregor, D. G. 2000. "Violence Risk Assessment and Risk Communication: The Effects of Using Actual Cases, Providing Instruction, and Employing Probability Versus Frequency Formats," *Law and human behavior* (24:3), pp. 271-296.
Spears, J. L., and Barki, H. 2010. "User Participation in Information Systems Security Risk Management," *MIS quarterly*), pp. 503-522.
Steelman, Z. R., Hammer, B. I., and Limayem, M. 2014. "Data Collection in the Digital Age: Innovative Alternatives to Student Samples," *MIS Quarterly* (38:2), pp. 355-378.
Steinbart, P. J., Keith, M. J., and Babb, J. 2016. "Examining the Continuance of Secure Behavior: A Longitudinal Field Study of Mobile Device Authentication," *Information Systems Research* (27:2), pp. 219-239.
Taha, M. M., Alhaj, T. A., Moktar, A. E., Salim, A. H., and Abdullah, S. M. 2013. "On Password Strength Measurements: Password Entropy and Password Quality," *2013 INTERNATIONAL CONFERENCE ON COMPUTING, ELECTRICAL AND ELECTRONIC ENGINEERING (ICCEEE)*: IEEE, pp. 497-501.
Tang, L. R., Jang, S. S., and Morrison, A. 2012. "Dual-Route Communication of Destination Websites," *Tourism Management* (33:1), pp. 38-49.
Ur, B., Kelley, P. G., Komanduri, S., Lee, J., Maass, M., Mazurek, M. L., Passaro, T., Shay, R., Vidas, T., and Bauer, L. 2012. "How Does Your Password Measure Up? The Effect of Strength Meters on Password Creation," *Presented as part of the 21st {USENIX} Security Symposium ({USENIX} Security 12)*, pp. 65-80.
Ur, B., Segreti, S. M., Bauer, L., Christin, N., Cranor, L. F., Komanduri, S., Kurilova, D., Mazurek, M. L., Melicher, W., and Shay, R. 2015. "Measuring Real-World Accuracies and Biases in Modeling Password Guessability," *24th {USENIX} Security Symposium ({USENIX} Security 15)*, pp. 463-481.
Vance, A., Eargle, D., Ouimet, K., and Straub, D. 2013. "Enhancing Password Security through Interactive Fear Appeals: A Web-Based Field Experiment," *2013 46th Hawaii International Conference on System Sciences*: IEEE, pp. 2988-2997.
Venkatesh, V., Brown, S. A., and Bala, H. 2013. "Bridging the Qualitative-Quantitative Divide: Guidelines for Conducting Mixed Methods Research in Information Systems," *MIS quarterly*), pp. 21-54.
Veras, R., Collins, C., and Thorpe, J. 2014. "On Semantic Patterns of Passwords and Their Security Impact," *NDSS*.
Verizon. 2017. "Data Breach Investigation Report: Understand Your Cybersecurity Threats." Retrieved August 15, 2017, from http://www.verizonenterprise.com/verizon-insights-lab/dbir/2017/
Visschers, V. H., Meertens, R. M., Passchier, W. W., and De Vries, N. N. 2009. "Probability Information in Risk Communication: A Review of the Research Literature," *Risk Analysis: An International Journal* (29:2), pp. 267-287.
Wei, W. 2017. "9 Popular Password Manager Apps Found Leaking Your Secrets." Retrieved August 15, 2017, from http://thehackernews.com/2017/02/password-manager-apps.html
Weir, M., Aggarwal, S., Collins, M., and Stern, H. 2010. "Testing Metrics for Password Creation Policies by Attacking Large Sets of Revealed Passwords," *Proceedings of the 17th ACM conference on Computer and communications security*, pp. 162-175.
Weir, M., Aggarwal, S., De Medeiros, B., and Glodek, B. 2009. "Password Cracking Using Probabilistic Context-Free Grammars," *2009 30th IEEE Symposium on Security and Privacy*: IEEE, pp. 391-405.
Yang, W., Li, N., Molloy, I. M., Park, Y., and Chari, S. N. 2016. "Comparing Password Ranking Algorithms on Real-World Password Datasets," *European Symposium on Research in Computer Security*: Springer, pp. 69-90.
Yang, Y., Yeo, K. C., Azam, S., Karim, A., Ahammad, R., and Mahmud, R. 2020. "Empirical Study of Password Strength Meter Design," *2020 5th International Conference on Communication and Electronics Systems (ICCES)*: IEEE, pp. 436-442.
Zhang, J., Luo, X., Akkaladevi, S., and Ziegelmayer, J. 2009. "Improving Multiple-Password Recall: An Empirical Study," *European Journal of Information Systems* (18:2), pp. 165-176.
Zviran, M., and Haga, W. J. 1999. "Password Security: An Empirical Study," *Journal of Management Information Systems* (15:4), pp. 161-185.

Online Appendix to

Augmenting Password Strength Meter Design using the Elaboration Likelihood Model: Evidence from Randomized Experiments

A. Markov Model and Markov Model with Backoff

A.1 Markov Model

N-gram models, i.e., Markov chains, have been applied to passwords (Castelluccia et al., 2012). A Markov chain of order d, where d is a positive integer, is a process that satisfies

$$P(x_i|x_{i-1}, x_{i-2}, \dots, x_1) = P(x_i|x_{i-1}, \dots, x_{i-d}),$$

where d is finite and x_1, x_2, x_3, . . . is a sequence of random variables. A Markov chain with order d corresponds to an n-gram model with$n = d + 1$.

For example, a 2-order Markov chain, in which the probability assigned to a password "$c_1 c_2 \cdots c_l$" is

$$P(c_1 c_2 \cdots c_l) = P(c_1|c_0)(c_2|c_1)(c_3|c_2) \cdots (c_l|c_{l-1}),$$

where $c_0 \notin \sum$ denotes a start symbol that is prepended to the beginning of all passwords, and

$$P(c_i|c_{i-1}) = \frac{count(c_{i-1}c_i)}{count(c_{i-1}\,\cdot)},$$

where $count(c_{i-1}\,\cdot)$ denotes the number of occurrences of c_{i-1} where it is followed by another character (i.e., where it is not at the end of password), and $count(c_{i-1}c_i)$ gives the number of occurrences of the substring $c_{i-1}c_i$. By definition, we have $count(c_{i-1}\,\cdot) = \sum_{c_i \in \sum} count(c_{i-1}c_i)$.

A.2 Markov Model with Backoff

Ma et al. (2014) proposed to use the Markov Model with backoff to model passwords. The intuition is that if a history appears frequently, then we would want to use that to estimate the probability of the next character. For example, if the prefix is *passwor*, using the whole prefix to estimate the next character would be more accurate than using only *wor*. On the other hand, if we observe a prefix that rarely appears, e.g., *!W}ab*, using only *ab* as the prefix is likely to be more accurate than using longer history, which likely would assign equal probability to all characters. One way to do this is Katz's backoff model (Katz 1987). In this model, one chooses a threshold and stores all substrings whose counts are above the threshold and use the frequency of these substrings to compute the probability.

B. Study 1 Questionnaire and Evaluation

B.1. Instructions and Survey Measures

Imagine you need to create a new account and password to access an online website such as a discussion forum. For example, craigslist, Reddit, or any other forum that you find interesting. To assist with account creation, the website has implemented a password strength meter and you have entered a password into the system. The calculated strength of the password you entered is indicated by the message to the right of the password.

For the purposes of this survey, please keep this image of the password strength meter in mind. [randomly assigned treatment screenshot shown to participants]

1. Optimistic Bias: Your Perceived Vulnerability

• How likely are you to become a victim of a stolen password?

• How likely is it that your personal information will be stolen over the Internet?

2. Optimistic Bias: Others' Perceived Vulnerability

• How likely are others to become a victim of a stolen password?

• How likely is it that others personal information will be stolen over the Internet?

3. Perceived Severity

• If my password was stolen, the consequences would be severe.

• If my password was stolen, the consequences would be serious.

• If my password was stolen, the consequences would be significant.

4. Perceived Susceptibility

• My password is at risk of being stolen.

• It is likely that my password will be stolen.

• It is possible that my password will be stolen. (item dropped from analysis)

5. Response Efficacy

• Using a hard-to-guess password works for protection. (reversed item)

• Using a hard-to-guess password is effective for protection. (reversed item)

• By using a hard-to-guess password, my password is more likely to be protected. (reversed item)

6. Self Efficacy

• Using a hard-to-guess password is easy to do. (reversed item)

• Using a hard-to-guess password is convenient to do. (reversed item)

• I am able to use a hard-to-guess password without much effort. (reversed item)

B.2 Measurement and Structural Model

Multivariate tests for reliability and validity of our data and measures are described below. Our examination of the loadings and cross-loadings of measures suggested that we drop one of the Perceived Susceptibility

measures. All the other measures for each construct were highly interrelated. Loadings are higher on the theoretically assigned constructs than the cross-loadings on the other constructs as shown by the bold-faced values in Table B1. All of the *p*-values for the outer model loadings are significant ($p < 0.01$) and establish convergent validity in our measures (Gefen and Straub 2005).

Table B1: Loadings and Cross-Loadings

Construct	Items	Perceived Severity	Perceived Susceptibility	Response Efficacy	Self Efficacy	Your Perceived Vul.	Others' Perceived Vul.
Perceived Severity	Seve1	**0.93**	0.12	-0.05	-0.12	0.04	0.06
	Seve2	**0.97**	0.14	-0.09	-0.08	0.04	0.06
	Seve3	**0.94**	0.12	-0.12	-0.07	0.067	0.07
Perceived Susceptibility	Susc1	0.16	**0.94**	0.07	0.19	0.61	0.07
	Susc2	0.08	**0.89**	0.02	0.16	0.68	0.18
Response Efficacy	Resp1	-0.08	-0.01	**0.88**	0.21	-0.02	-0.02
	Resp2	-0.07	0.08	**0.93**	0.17	-0.01	-0.06
	Resp3	-0.10	0.04	**0.91**	0.17	0.01	-0.04
Self Efficacy	Self1	-0.08	0.13	0.21	**0.81**	0.19	0.10
	Self2	-0.07	0.05	0.11	**0.71**	0.09	0.12
	Self3	-0.08	0.23	0.15	**0.91**	0.27	0.11
Your Perceived Vulnerability	Your1	0.02	0.66	-0.02	0.26	**0.95**	0.18
	Your2	0.09	0.62	0.02	0.19	**0.90**	0.25
Others' Perceived Vulnerability	Othr1	0.06	0.11	-0.05	0.14	0.20	**0.96**
	Othr2	0.08	0.12	-0.03	0.07	0.24	**0.85**

Regarding discriminant validity (shown in Table B2), the square root of the average variance extracted (AVE) for each construct is much larger than any correlation among the other constructs. Also, the lack of cross-loading provides sufficient evidence of discriminant validity (Gefen and Straub 2005). Further, the composite reliability and values for Cronbach's α are higher than the threshold recommended by the literature (Chin 1998; Fornell and Larcker 1981), indicating internal consistency. In summary, our measures appear to be internally consistent and demonstrate convergent and discriminant reliability.

Table B2: Reliability and Interconstruct Correlations

Construct	Composite Reliability	Cronbach's *a*	Perceived Severity	Perceived Susc.	Resp. Efficacy	Self Efficacy	Your Perceived Vul.	Others' Perceived Vul.
Perceived Sev.	0.96	0.94	**0.95**					
Perceived Susc.	0.91	0.82	0.14	**0.92**				
Resp. Efficacy	0.93	0.90	-0.09	0.05	**0.91**			
Self Efficacy	0.85	0.76	-0.10	0.19	0.20	**0.81**		
Your Perceived Vul.	0.92	0.83	0.05	0.69	-0.00	0.25	**0.92**	
Others' Perceived Vul.	0.90	0.80	0.07	0.13	-0.05	0.13	0.23	**0.91**
Notes. Values in bold are the square root of the average variance extracted								

B2. Results

We present the results of the PLS analyses for our exploratory experiment in Table B3. It considers all the data in a combined analysis (i.e., 253 observations including all treatments: control, time, rank, probability). We used SmartPLS 3.2.7 (Ringle et al. 2015) for the analyses and specified bootstrapping with 500 re-samples for calculating the standard errors. Similar to the literature, we observe significant coefficients for most of the relationships, with just a few exceptions as noted in the table.

Table B3: Summary of Path Coefficients

Relationship	β	**Significance**
Perceived Severity ↔ Response Efficacy	-0.10	$p < 0.10$
Perceived Severity ↔ Self Efficacy	-0.12	$p < 0.05$
Perceived Susceptibility ↔ Response Efficacy	0.07	-
Perceived Susceptibility ↔ Self Efficacy	0.21	$p < 0.01$
Response Efficacy ↔ Your Perceived Vulnerability	-0.06	-
Response Efficacy ↔ Others' Perceived Vulnerability	-0.07	-
Self Efficacy ↔ Your Perceived Vulnerability	0.26	$p < 0.01$
Self Efficacy ↔ Others' Perceived Vulnerability	0.14	$p < 0.05$

C. Additional Information of Study 2

C.1 Instructions

You are participating in a study regarding password generation and usage. Please be aware that passwords you generate in this study will be recorded.

• Please DO NOT use any of your existing password(s).

• However, as the goal of our study is to understand how users choose passwords, we would appreciate it if you use your regular password generation strategy to generate passwords in the study.

C.2 Scenarios

Each participant is randomly assigned to one of the following three scenarios.

• A bank named "Provident Bank" just opened a new branch close to your home. You have opened a checking account and obtained a credit card with them. You deposited $10,000 into that checking account and the bank gave you a $5,000 credit line for your credit card. Now you are in the process of creating an online banking account.

• You are ordering food online at a website called "Food Express". During the ordering process, you have entered your name, delivery address, and phone number, which will be stored on the website. Now at the end of the ordering process, the website requires you to create an account with them.

• You and your friends are going to Disney World. You want to find more details regarding one specific attraction and would like to ask about it online. You found a website called "Disney World Friendly Forum", and decided to post a question on that forum. In order to do so, you have to create an account first. Only your email address and a password are needed.

C.3 Password Confirmation

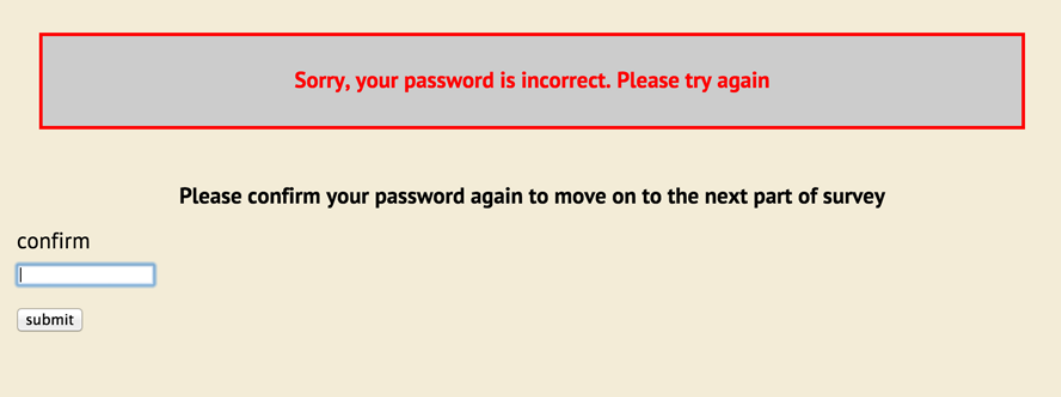

Figure C1: Password Confirmation Page

C.4 Demographic Survey

Please answer the following questions.

1. How old are you?

• Under 18

• 18-22

• 23-30

• 31-40

• 41-50

• 51+

• Would rather not say

2. What is your gender?

• Male

• Female

• Would rather not say

3. What is the highest degree or level of school you have completed?

• High school graduate.

• Some college credit, no degree.

• Bachelor's degree.

• Master's degree.

• Professional degree.

• Doctorate degree.

• Would rather not say

C.5. Total number of participants in each treatment/scenario

Table C1: Total number of participants in each treatment/scenario

	Time	**Rank**	**Probability**	**Control**	**Total**
Bank	41	55	47	41	184
Restaurant	51	36	43	36	166
Forum	41	40	30	39	150
Total	133	131	120	116	500

C.6. Post-Test Survey

In a laboratory experiment, it is important that we implement the study in such a way that potential concerns about participants entering unusable passwords or not taking the task seriously can be sufficiently addressed. Therefore, we conducted a post-test survey as suggested by Goodman et al. (2013) to improve confidence regarding the reliability of participants' data. The post-test survey was conducted as follows. After a participant submits their password, they are asked to participate in a survey focused on their password generation strategy. Note that the survey requests information about how the participant generates their passwords in general, not only the password they just created for this study. The post-test survey is

intentionally non-trivial, as it takes approximately 30 minutes to finish. After completing the survey, the system then asks participants to re-enter the password created in the initial password strength meter task. Following that, regardless of whether the participant can successfully recall the password she generated or not, demographic questions such as age, gender, and education level, are captured for the purpose of statistical controls. We intended to exclude any participant who failed to properly re-enter their previously generated password from the experiment. However, we found that no participant failed the post-test survey reliability check.

D. Additional Information of Study 3

In Study 3, we utilize the same contextual message used in the laboratory experiment (i.e., Study 2). Nevertheless, the appearance is slightly different. Figures below show the screenshot of each type of password strength meter used in the experiment.

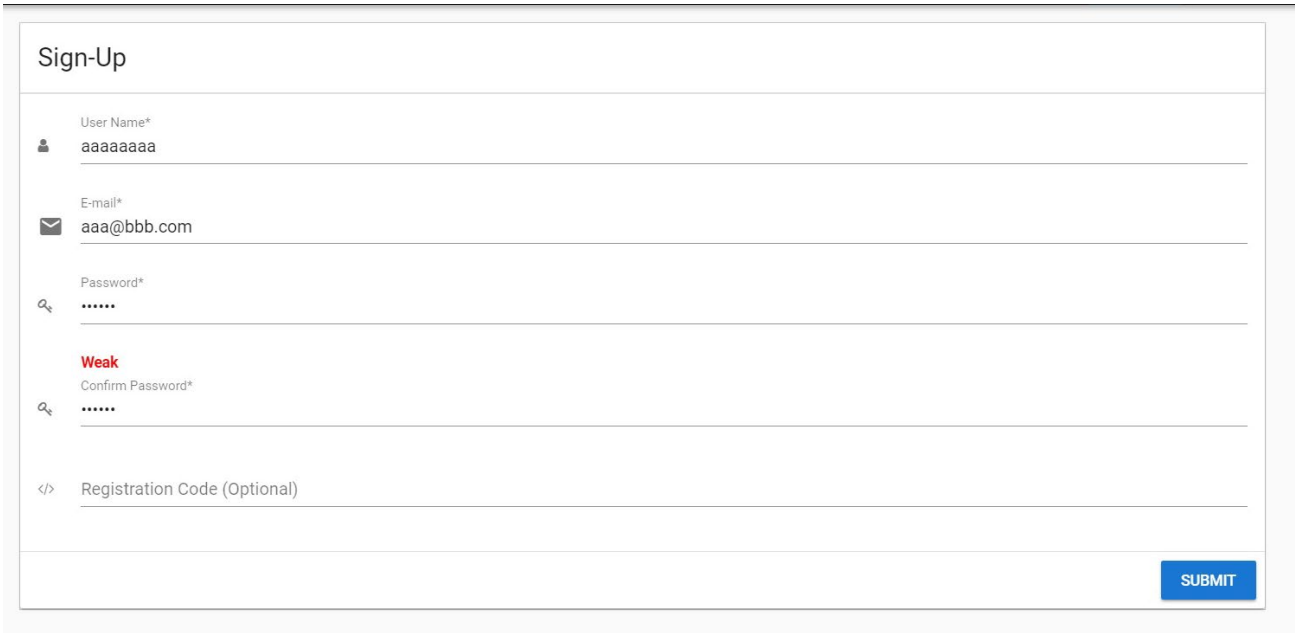

Figure D1: The password strength meter under *Control* treatment

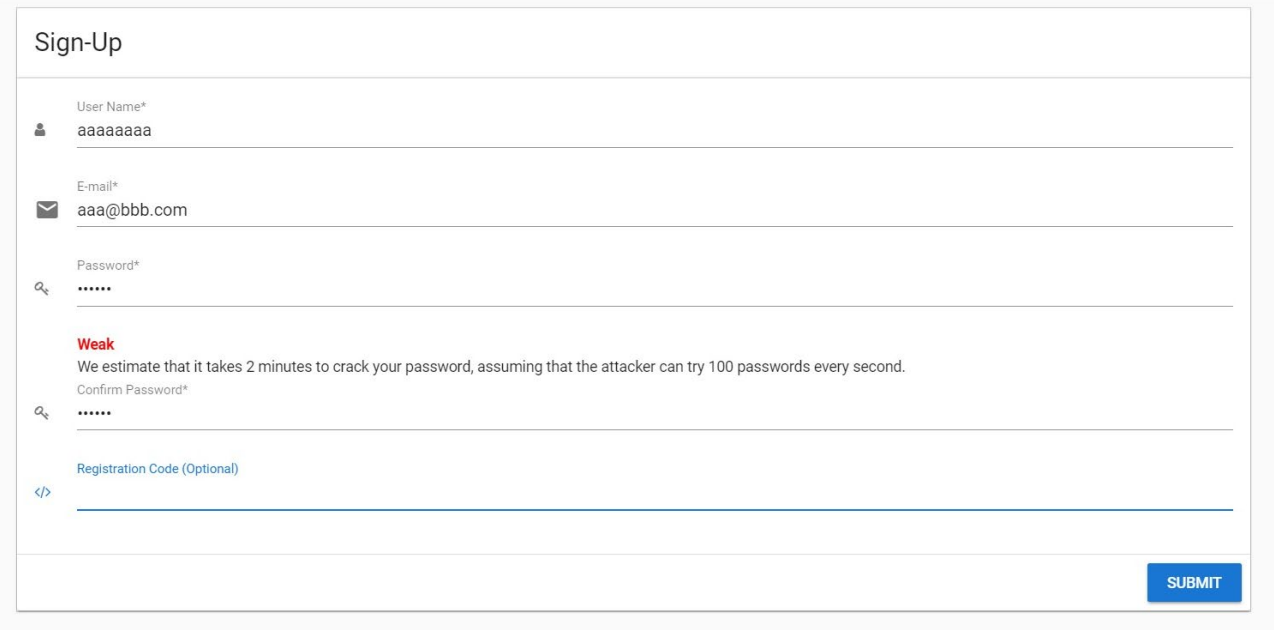

Figure D2: The password strength meter under *Time* treatment

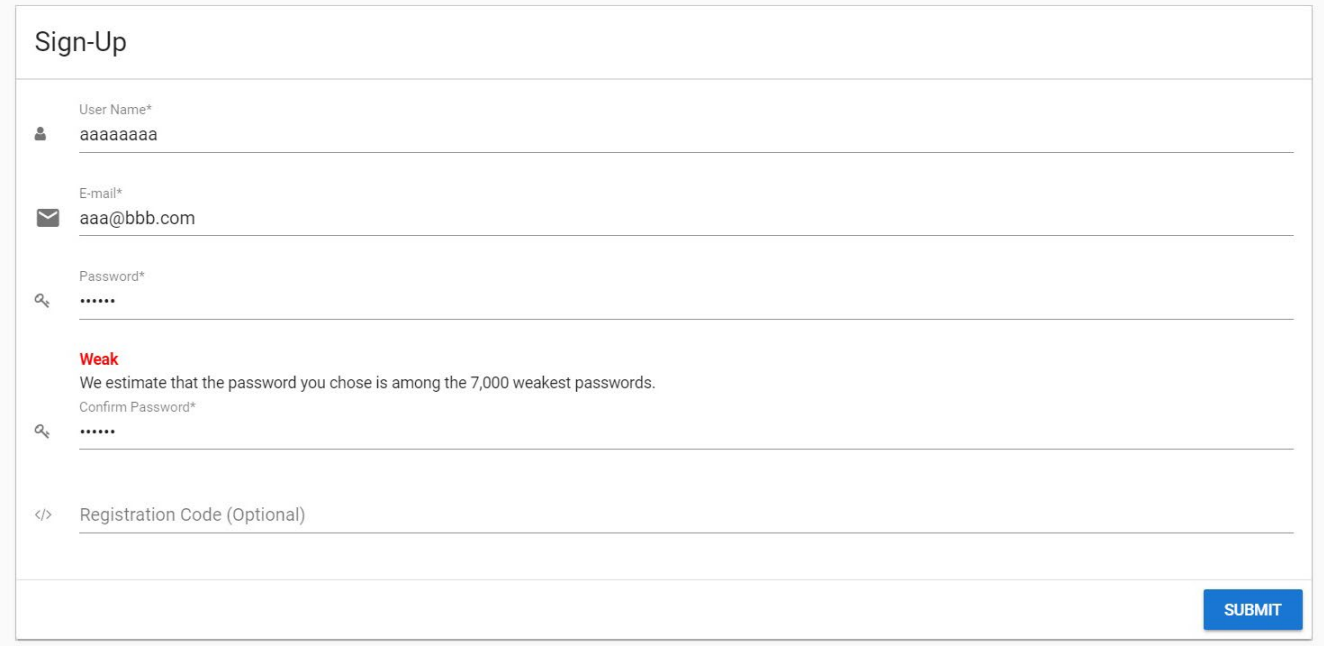

Figure D3: The password strength meter under *Rank* treatment

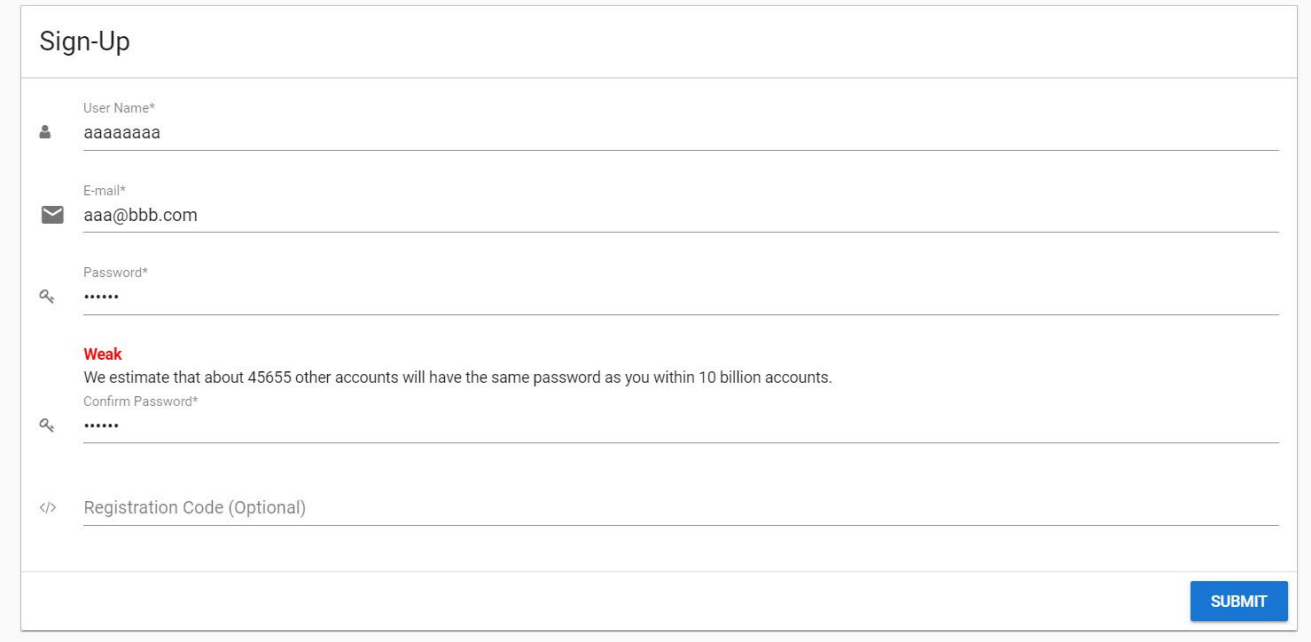

Figure D4: The password strength meter under *Probability* treatment

E. Additional Analyses

This appendix reports the results of additional analyses we conduct to enhance the robustness of our results.

E.1 Controlling for the Effort to Learn More about Password Security in Study 2

In this appendix, we add the variable *learn_more* as a control variable in the regression analyses that examine *diff_strength* and *num_reset*. All other specifications remain similar to those in the main models.

Table E1-1: Results from the regression analysis

	diff_strength	*num_reset*
Treatment		
time	0.034 (0.028)	0.452 (0.358)
rank	**0.091** (0.028)**	**1.075** (0.322)**
probability	0.048 (0.029)	0.107 (0.339)
Scenario		
restaurant	0.047 (0.024)	0.366 (0.252)
forum	0.064** (0.024)	0.504 (0.256)
Control		
age	-0.001 (0.010)	-0.050 (0.100)
gender	0.011 (0.020)	-0.161 (0.210)
education	0.002 (0.011)	-0.197 (0.128)
learn_more	0.037 (0.029)	1.822*** (0.204)

Constant	-0.045 (0.048)	-2.371*** (0.534)
Observations	497	497
R-squared	0.037	0.156
Notes. Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$		

Table E1-2: Results from the regression analysis with initial strength

	diff_strength
Treatment	
time	0.033 (0.027)
rank	**0.090** (0.027)**
probability	0.045 (0.028)
Scenario	
restaurant	0.039 (0.023)
forum	0.046 (0.024)
Control	
age	0.004 (0.009)
gender	0.009 (0.019)
education	0.002 (0.010)
Initial strength	-0.087*** (0.014)
learn_more	0.042 (0.028)
Constant	0.177** (0.059)
Observations	497
R-squared	0.107
Notes. Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$	

E.2 The Mediation Effects of *num_reset* in Study 2

We follow the approach used in Baron and Kenny (1986) to test the mediator effect of *num_reset* on the effect of context-based password strength meters on the increased in the strength of the passwords generated that we examine in our Study 2.

In step 1, we regress the dependent variable (*diff_strength*) on the independent variable (*Time*, *Rank*, and *Probability*). The purpose of this step is to verify that the independent variable is a predictor of the dependent variable. The results are reported in Table E2-1 below.

Table E2-1: Step one of the mediator analysis

	diff_strength
Treatment	
time	0.031 (0.028)
rank	**0.089** (0.028)**
probability	0.046 (0.029)

Scenario	
restaurant	0.047 (0.024)
forum	0.063* (0.024)
Control	
age	-0.001 (0.010)
gender	0.011 (0.020)
education	0.002 (0.011)
Constant	-0.039 (0.048)
Observations	497
R-squared	0.034
Notes. Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$	

In step 2, we regress the mediator (*num_reset*) on the independent variable variable (*Time*, *Rank*, and *Probability*) to confirm that the independent variable is a predictor of the mediator. The results are reported in Table E2-2 below.

Table E2-2: Step two of the mediator analysis

	num_reset
Treatment	
time	0.288 (0.357)
rank	**0.993** (0.322)**
probability	0.107 (0.339)
Scenario	
restaurant	0.370 (0.251)
forum	0.389 (0.255)
Control	
age	-0.079 (0.100)
gender	-0.119 (0.209)
education	-0.147 (0.119)
Constant	-1.754*** (0.524)
Observations	497
R-squared	0.033
Notes. Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$	

In step 3, we regress the dependent variable (*diff_strength*) on both the mediator (*num_reset*) and independent variable (*Time*, *Rank*, and *Probability*) to confirm that the mediator is a predictor of the dependent variable. The results are reported in Table E2-3 below. We observe that the coefficients of all treatments are lower,

and the statistical significance of *Rank* changes from $p < 0.01$ in step 1 to $p < 0.05$ in step 2. Meanwhile, the coefficient of *num_reset* is still statistically significant at $p < 0.001$. Hence, we can conclude that *num_reset* is a mediator on the effect of context-based password strength meters on the increased in the strength of the passwords generated.

Table E2-3: Step three of the mediator analysis

	diff_strength
Treatment	
time	0.027 (0.027)
rank	**0.069* (0.027)**
probability	0.034 (0.028)
Scenario	
restaurant	0.039 (0.023)
forum	0.055* (0.023)
Control	
age	0.001 (0.009)
gender	0.013 (0.019)
education	0.005 (0.010)
num_reset	0.109*** (0.016)
Constant	-0.060 (0.047)
Observations	497
R-squared	0.115
Notes. Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$	

E.3 The Moderation Effects of *initial_strength_label* in Study 2 and Study 3

We consider *initial_strength_label* as a moderator for the effect of context-based password strength meters on password-generating behavior. The results in Table E3-1 below report the moderation effect of *initial_strength_label* based on the data from the laboratory experiment. In the meantime, Table E3-2 reports the moderation effect of *initial_strength_label* on *diff_strength* and Table E3-3 reports the moderation effect of *initial_strength_label* on *num_reset* based on the data from the field experiment. Most of the results in this extension remain qualitatively similar to the main results. The only exception is that the treatment effect on *num_reset* becomes statistically insignificant at $p < 0.05$ once the *initial_strength_label* is added as a moderator. As for the moderation effect, both results from Study 2 and Study 3 demonstrate that *initial_strength_label* is a strong moderator for *diff_strength* but not a moderator for *num_reset*.

Table E3-1: The moderation analysis of *initial_strength_label* from Study 2

	diff_strength	*num_reset*	*learn_more*
Treatment			
time	0.023 (0.129)	1.041 (1.217)	-3.893 (2.049)
time x *initial_strength_label*	0.003 (0.040)	-0.308 (0.469)	1.251 (0.723)
rank	**0.389*** (0.110)**	1.424 (1.142)	-1.884 (1.488)
rank x *initial_strength_label*	**-0.114** (0.041)**	-0.1775 (0.436)	0.618 (0.546)
probability	0.039 (0.109)	1.484 (1.149)	-2.293 (1.573)
probability x *initial_strength_label*	0.002 (0.040)	-0.333 (0.443)	0.719 (0.574)
Scenario			
restaurant	0.033 (0.023)	0.312 (0.253)	-0.004 (0.309)
forum	0.042 (0.024)	0.309 (0.260)	-0.345 (0.341)
Control			
age	0.003 (0.009)	-0.050 (0.101)	-0.001 (0.129)
gender	0.011 (0.019)	-0.103 (0.210)	-0.083 (0.269)
education	0.002 (0.010)	-0.156 (0.121)	0.043 (0.142)
initial_strength_label	-0.060* (0.030)	-0.167 (0.3843)	-0.394 (0.346)
Constant	0.115 (0.093)	-1.352 (1.132)	-0.510 (1.099)
Observations	497	497	497
R-squared	0.126	0.051	0.020
Notes. Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$			

Table E3-2: The moderation analysis of *initial_strength_label* on *diff_strength* from Study 3

	Sum of squares	**d.f.**	**Mean square**	***F*-ratio**	**Significance**
Model	9.970	11	0.906	9.13	<0.001
Treatment	1.865	3	0.622	6.26	<0.001
initial_strength_label	5.978	2	2.990	30.11	<0.001
Treatment × *initial_strength_label*	1.268	6	0.211	2.13	0.050
Residual	29.383	296	0.993		
Total	39.353	307	0.128		

Table E3-3: The moderation analysis of *initial_strength_label* on *num_reset* from Study 3

	Sum of squares	**d.f.**	**Mean square**	***F*-ratio**	**Significance**
Model	15.762	11	1.433	2.90	0.001
Treatment	7.806	3	2.602	5.27	0.001
initial_strength_label	5.274	2	2.637	5.34	0.005
Treatment × *initial_strength_label*	1.924	6	0.321	0.65	0.690
Residual	146.066	296	0.493		
Total	161.828	307	0.527		

E.4 Alternative Password Strength Calculation in Study 2

In the paper, we calculate password strength using the Backoff Markov Model with end symbol normalization proposed by Ma et al. (2014). Since the password strength generates by this approach is stochastic (i.e., can be changed based on the training dataset), we consider an alternative password strength calculation where the process is deterministic. Here, we consider the password entropy, which is commonly used in the literature (e.g., Houshmand and Aggarwal 2012; Weir et al. 2010), as an alternative password strength calculation. The results are reported in Table E4-2 below.

Table E4-1: Alternative Password Strength Calculation (Password Entropy)

	diff_strength
Treatment	
time	0.031 (0.019)
rank	**0.057** (0.020)**
probability	**0.040* (0.020)**
Scenario	
restaurant	0.022 (0.016)
forum	0.042* (0.017)
Control	
age	-0.001 (0.007)
gender	-0.010 (0.014)
education	0.001 (0.007)
Constant	-0.024 (0.033)
Observations	497
R-squared	0.030
Notes. Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$	

E.5 Randomization Checks in Study 1 and Study 2

In Study 1 and Study 2, we conduct a randomized experiment where participants are randomly assigned to one of the treatments. In this section, we perform randomization checks to show that the different treatment groups, in the same experiment, are not systematically different from one other. Specifically, we list descriptive statistics for different treatment groups in Study 1 and Study 2 and perform ANOVA test to verify if any systematic difference exists. Results show no systematic difference in all variables used in both Study 1 and Study 2.

Table E5-1: Descriptive Statistics of Participants in Study 1 (Mean and Standard Deviation are Reported)

	Control	Time	Rank	Probability
Gender	0.600 (0.494)	0.656 (0.479)	0.547 (0.502)	0.508 (0.504)
Age	21.600 (2.866)	21.689 (2.248)	21.563 (3.528)	21.524 (1.401)
n	65	61	64	63

Table E5-2: Randomization Checks for Gender in Study 1

	Sum of squares	**d.f.**	**Mean square**	***F*-ratio**	**Significance**
Between Groups	0.771	3	0.257	1.05	0.371
Within Groups	60.976	249	0.245		
Total	61.747	252	0.245		

Table E5-3: Randomization Checks for Age in Study 1

	Sum of squares	**d.f.**	**Mean square**	***F*-ratio**	**Significance**
Between Groups	0.921	3	0.307	0.04	0.988
Within Groups	1730.146	249	6.948		
Total	1731.067	252	6.869		

Table E5-4: Descriptive Statistics of Participants in Study 2 (Mean and Standard Deviation are Reported)

	Control	Time	Rank	Probability
Gender	0.483 (0.502)	0.429 (0.497)	0.473 (0.531)	0.375 (0.503)
Age	3.448 (0.999)	3.624 (1.105)	3.595 (1.115)	3.475 (1.061)
Education	2.638 (0.955)	2.654 (1.008)	2.626 (0.914)	2.600 (0.911)
n	116	133	131	120

Table E5-5: Randomization Checks for Gender in Study 2

	Sum of squares	**d.f.**	**Mean square**	***F*-ratio**	**Significance**
Between Groups	0.882	3	0.294	1.14	0.334
Within Groups	128.318	496	0.259		
Total	129.200	499	0.259		

Table E5-6: Randomization Checks for Age in Study 2

	Sum of squares	**d.f.**	**Mean square**	***F*-ratio**	**Significance**
Between Groups	2.825	3	0.942	0.82	0.485
Within Groups	571.375	496	1.152		
Total	574.200	499	1.151		

Table E5-7: Randomization Checks for Education in Study 2

	Sum of squares	**d.f.**	**Mean square**	***F*-ratio**	**Significance**
Between Groups	0.195	3	0.650	0.07	0.976
Within Groups	464.355	496	0.936		
Total	464.550	499	0.931		

E.6 Controlling for *initial_strength* in Study 3

Similar to Study 2, we control for *initial_strength* in the analysis of Study 3. The results are reported in Table E6-1. They are consistent with results in the main paper.

Table E6-1: Controlling for *initial_strength* in Study 3

	diff_strength
Treatment	
time	**0.111* (0.052)**
rank	**0.147** (0.053)**
probability	**0.097* (0.052)**
Control	
initial_strength	-0.201*** (0.265)
Constant	0.506 (0.074)
Observations	308
R-squared	0.181
Notes. Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$	

REFERENCES

Baron, R. M., and Kenny, D. A. 1986. "The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of personality and social psychology* (51:6), p. 1173.

Chin, W. 1998. "Issues and Opinion on Structural Equation Modeling," *MIS Quarterly* (22:1), pp. vii-vii.

Fornell, C., and Larcker, D. F. 1981. "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of marketing research* (18:1), pp. 39-50.

Gefen, D., and Straub, D. 2005. "A Practical Guide to Factorial Validity Using Pls-Graph: Tutorial and Annotated Example," *Communications of the Association for Information systems* (16:1), p. 5.

Goodman, J. K., Cryder, C. E., and Cheema, A. 2013. "Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples," *Journal of Behavioral Decision Making* (26:3), pp. 213-224.

Houshmand, S., and Aggarwal, S. 2012. "Building Better Passwords Using Probabilistic Techniques," *Proceedings of the 28th Annual Computer Security Applications Conference*, pp. 109-118.

Katz, S. 1987. "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE transactions on acoustics, speech, and signal processing* (35:3), pp. 400-401.

Ma, J., Yang, W., Luo, M., and Li, N. 2014. "A Study of Probabilistic Password Models," *2014 IEEE Symposium on Security and Privacy*: IEEE, pp. 689-704.

Ringle, C. M., Wende, S., and Becker, J. 2015. "Smartpls 3. Bönningstedt: Smartpls, 2015."

Weir, M., Aggarwal, S., Collins, M., and Stern, H. 2010. "Testing Metrics for Password Creation Policies by Attacking Large Sets of Revealed Passwords," *Proceedings of the 17th ACM conference on Computer and communications security*, pp. 162-175.