Computers in Human Behavior 109 (2020) 106347

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: http://www.elsevier.com/locate/comphumbeh

Full length article

Nudge me right: Personalizing online security nudges to people's decision-making styles

Eyal Peer [a,*], Serge Egelman [b,c], Marian Harbach [b], Nathan Malkin [c], Arunesh Mathur [d], Alisa Frik [b,c]

[a] Federmann School of Public Policy, Hebrew University of Jerusalem, Israel
[b] International Computer Science Institute, Berkeley, CA, USA
[c] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA
[d] Department of Computer Science, Princeton University, USA

A B S T R A C T

Nudges are simple and effective interventions that alter the architecture in which people make choices in order to help them make decisions that could benefit themselves or society. For many years, researchers and practitioners have used online nudges to encourage users to choose stronger and safer passwords. However, the effects of such nudges have been limited to local maxima, because they are designed with the "average" person in mind, instead of being customized to different individuals. We present a novel approach that analyzes individual differences in traits of decision-making style and, based on this analysis, selects which, from an array of online password nudges, would be the most effective nudge each user should receive. In two large-scale online studies, we show that such personalized nudges can lead to considerably better outcomes, increasing nudges' effectiveness up to four times compared to administering "one-size-fits-all" nudges. We regard these novel findings a proof-of-concept that should steer more researchers, practitioners and policy-makers to develop and apply more efforts that could guarantee that each user is nudged in a way most right for them.

1. Nudge Me right: personalizing online security nudges to people's decision-making styles

When it comes to improving people's decisions, nudges—changes in choice architecture that predictably influence decisions (Thaler & Sunstein, 2008) —have been shown to be effective across important domains such as finance (e.g., Cai, 2019), health (Quigley, 2013), education (Damgaard & Nielsen, 2018), ethics (e.g., Bazerman & Gino, 2012), privacy and security (Acquisti et al., 2017; Collier, 2018) and more (e.g., Szaszi, Palinkas, Palfi, Szollosi, & Aczel, 2018). A nudge is any aspect of the choice architecture that alters people's behavior without a) forbidding or adding any relevant options or b) significantly changing their economic incentives (Hansen, 2016). Prominent examples include using defaults to auto-enroll people to pension plans (Carroll, Choi, Laibson, Madrian, & Metrick, 2009) or to register for organ donation (Johnson & Goldstein, 2003); adding traffic-light labels to food products to encourage healthier eating (VanEpps, Downs, & Loewenstein, 2016); asking people for pre-commitments to increase their future savings (Thaler & Benartzi, 2004) or their charitable giving (Breman, 2011); using prompts and reminders to increase vaccination rates (Milkman, Beshears, Choi, Laibson, & Madrian, 2011) or savings (Karlan, McConnell, Mullainathan, & Zinman, 2016); presenting fuel and energy efficiency as relative costs to encourage pro-environmental choices (Camilleri & Larrick, 2014); comparing households' electricity consumption to social norms to curb energy consumption (Costa & Kahn, 2013), and many more. Such, and many other nudges are now being effectively used at scale by many governments and public institutions (Organisation de Coopérationet de & Développement Economiques, 2017), and garner considerable support from the public in many countries (e.g., Reisch & Sunstein, 2016). The notion of nudging argues that, instead of relying on strict policies of mandates, bans, and sanctions, one can sometimes achieve better results if the desired choice is made easier, more attractive, socially desirable and given at the correct timing (The Behavioral Insights Team, 2014).

In the past years, researchers in the computer security community have been exploring nudges to promote better online security decision-making, such as using website privacy policy indicators, password strength meters, and others (Acquisti et al., 2017). Currently, though, computer security mitigations take a one-size-fits-all approach: every user sees the same messaging, regardless of their individual differences. However, the high variance among users (in risk aversion or technical proficiency, for example) and between situations (e.g., at home vs. in the

* Corresponding author. Hebrew University of Jerusalem, School of Public Policy, Mount Scopus, Jerusalem, 000000, Israel.
E-mail address: eyal.peer@mail.huji.ac.il (E. Peer).

https://doi.org/10.1016/j.chb.2020.106347
Received 13 November 2019; Received in revised form 25 February 2020; Accepted 15 March 2020
Available online 24 March 2020

workplace) is liable to weaken the efficacy of such an "average user" approach. No single security messaging has indeed been shown to deter most users from engaging in risky behavior. For example, compliance rates for security warnings are often quite low, despite significant progress being made (e.g., Akhawe & Felt, 2013; Reeder et al., 2018). Research has also found significant individual differences in computer security (Jeske, Coventry, Briggs, & van Moorsel, 2014; Malkin, Mathur, Harbach, & Egelman, 2017) and privacy (Egelman & Peer, 2015, pp. 16–28) behaviors, and has argued for the personalization of nudges in that domain.

The problem of nudges (or other interventions) being administered only according to averages in a "one-size-fits-all" approach is not limited to computer security. A particular nudge may have a strong positive effect on some individuals, but smaller, insignificant, or even negative effects on others, for whom a different nudge may be more effective. For example, Halpern (2016) describes a field experiment aimed to increase tax report rates using letters stressing social norms of neighborhoods' payment rates. While successful for most tax-payers, this approach caused a negative reaction among the top 5% of debtors, who held the largest debts (Halpern, 2016). Similarly, Costa and Kahn (2013) found that a social norms intervention to reduce electricity bills affected only liberal households (compared to conservatives). Thus, it appears that nudges' effects are confined to local maxima determined by the populations' heterogeneity.

In marketing, the emblematic solution to address consumer heterogeneity is typically personalization (see, e.g., Vesanen, 2007). Personalization includes a range of marketing techniques ranging from identifying customers' segmented needs to morphing a website based on cognitive style inferred from users' clickstream (Hauser, Urban, Liberali, & Braun, 2009), often at the expense of consumers' privacy (e.g., Chellappa & Sin, 2005). Because nudges are aimed to increase individuals' welfare, using personalization techniques to provide the best-fitting nudges to different people should be favored, from a managerial and policy perspective, whenever possible. Indeed, several scholars have already stressed the importance of targeting nudges in general (e.g., Goldstein, Johnson, Herrmann, & Heitmann, 2008; Sunstein, 2013). In the computer security domain, Shillair et al. (2015) also found significant interactions between users' traits (e.g., personal responsibility) and the type of intervention used to enhance users' online safety behavioral intentions. Based on that, they recommended that, in the future, intervention strategies should match the user's characteristics in order to maximize their potency. However, to date, only a few researchers tried to rise to that challenge and test such approaches experimentally.

An exception can be found in a study that showed how advertisements were evaluated more favorably if they were personalized to consumer's personality profile (Hirsh, Kang, & Bodenhausen, 2012). Namely, participants' purchasing intentions following an ad were found to be higher if the ad highlighted a personality trait on which they scored high. This may be the case when advertisements attempt to increase hypothetical purchasing intentions. Nevertheless, this approach requires tailoring the persuasive message to the target consumer's characteristics. Alternatively, it is possible that a stronger outcome could be achieved if existing nudges, which have already been shown to work on average, are deliberately given only to the specific groups of individuals on which they are expected, ex ante, to yield a positive effect, while other groups would receive different nudges or be treated differently. In other words, personalization could be more effective if it is directed at selecting a nudge from a pool of existing nudges. In this we offer to distinguish between *personalizing the nudge* (e.g., adding the recipient's first name to the nudge's message) vs. *personalizing which nudge* (e.g., assigning different kinds or versions of nudges to different individuals), and focus our study on the latter.

Testing this kind of nudge personalization could be possible in decision situations in which different nudges are sometimes used in parallel to achieve the same goal. One area in which nudges are used regularly and in diverse methods, and could provide the adequate testbed to explore nudge personalization, lies in the domain of cybersecurity and, specifically, the area of online computer security. Specifically, researchers and practitioners have developed various nudges that are aimed at encouraging users to create and use stronger passwords on their computers and online services. This specific, yet ubiquitous, area could be where personalized nudging may have great potential, and thus we focused our research on nudges aimed at encouraging online users to create more secure passwords. Based on the literature in this area, we focused on five password nudges that are either often used or frequently researched:

1) Password Meter: Visual display of real-time feedback regarding the password's strength (Ur et al., 2012). Egelman, Sotirakopoulos, Muslukhov, Beznosov, and Herley (2013) found that password meters can increase password strength by approximately 30% on average, compared to a control condition.
2) Crack-Time: Quantitative feedback of how long it would take a hacker to crack the password (Wheeler, 2016). Vance, Eargle, Ouimet, and Straub (2013) found that participants given this nudge selected stronger passwords than those who got a regular password meter.
3) Social: Comparing the password's strength to social norms (i.e., the strength of the user's password relative to others on the system) Egelman et al. (2013) found that this comparison can increase password strength by approximately 32% on average, compared to a control condition.
4) CHBS (Correct Horse Battery Staple): Suggesting that users create passwords by concatenating a series of words together. Shay et al. (2012) showed this nudge can significantly increase the strength of password people choose, as well as increase their degree of accurate recall after time.
5) Insertion: Suggesting that users randomly insert numbers and special characters into their chosen passwords. Forget, Chiasson, van Oorschot, and Biddle (2008) showed that giving this nudge to users can increase their password's strength by 10%–65%, depending on how many new characters are inserted, without it increasing the number of errors users made when trying to recall their password.

All of these password nudges are, basically, persuasion messages advising people how to choose better passwords. They can be regarded as nudges because they alter key aspects of the choice architecture in which users are asked to create their passwords online. Specifically, the first password meter and social nudges do so by making specific information available and salient, and present that information in a specific manner that affects users' decisions. Namely, they provide users with a better sense of how strong (or weak) the password they are creating is, and also provide a dynamic feedback on how changes in the password (e.g., adding a special character) can enhance it. Importantly though, these nudges do not restrict users' choices (like security policies do) and thus fit the basic definition of nudges (Thaler & Sunstein, 2008). The crack-time nudge goes one step beyond the regular meters and provides a novel measure of how strong the password is, also highlighting the risk of an attacker cracking a weak password quicker and easier. The other two nudges, CHBS and Insertion, work somewhat differently as they do not provide feedback on the user's chosen password but, rather they provide users with simple heuristics on how they could easily generate strong passwords. These heuristics try to prevent users from the habit of using memorable or previously used passwords that might be weaker, and they make use of people's tendency to follow simple heuristics (which the nudges provide). By this, CHBS and Insertion fit newer and more nuanced definitions of nudges (Hansen, 2016). Even though other password nudges exist (e.g., Renaud & Zimmermann, 2019), we focused on these selected sub-set of password nudges to examine how they could be personalized by assigning different users to different password nudges.

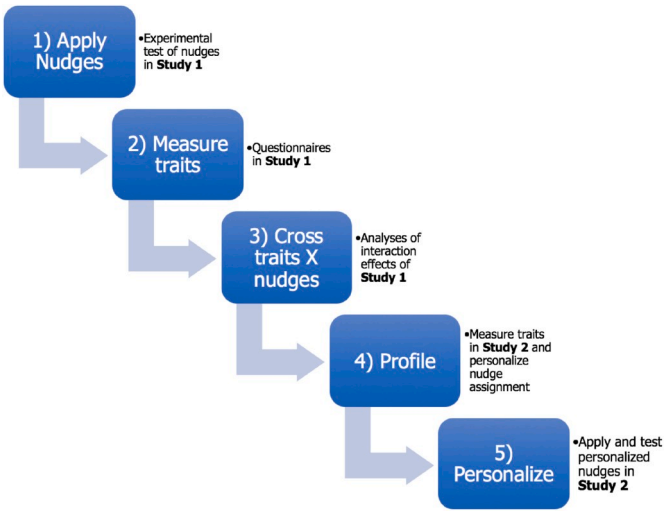


Fig. 1. Workflow of developing and testing nudge personalization, as applied in the studies.

Whereas previous research on personalizing persuasion messages focused on personality traits (e.g., Hirsh et al., 2012), we posited that when nudges are used to influence people's actual choices and decisions (and not mere intentions), a better fit could be achieved when personalizing the nudge according to individual traits that relate to decision making (e.g., Appelt, Milch, Handgraaf, & Weber, 2011; Egelman & Peer, 2015, pp. 16–28). We thus focused on individual differences in decision-making styles that can be measured using validated scales from the decision-making literature and have been found to moderate people's judgment and decision-making in certain situations, and were thus expected to be potential candidates for personalization.

1) General Decision-Making Style (GDMS, Scott & Bruce, 1995) - how individuals approach decision situations, including tendencies for *rational, avoidant, dependent, intuitive,* and *spontaneous* styles of decision-making, each measured on a sub-scale in the GDMS. The GDMS has been extensively used and was found, among other things, to predict the type of advice people prefer to receive (Dalal & Bonaccio, 2010).
2) Need for Cognition (NFC, Cacioppo, Petty, & Feng Kao, 1984) - people's tendency to engage in high thought-effort tasks. NFC was found to moderate reactions to framing effects (Smith & Levin, 1996) and message persuasion effects (Cacioppo, Petty, & Morris, 1983).
3) Consideration for Future Consequences (CFC, Strathman, Gleicher, Boninger, & Edwards, 1994) - the extent to which people consider distant versus immediate consequences of potential behaviors. CFC has been found to predict, for example, the extent to which people take on risk (Zimbardo, Keough, & Boyd, 1997), which could be highly relevant to decisions about cybersecurity and password strength.
4) Numeracy (Peters et al., 2006) - people's basic quantitative skills, as measured using simple arithmetic questions that have been extensively developed and researched. Numeracy is a known predictor of decisions under risk (e.g., Reyna, Nelson, Han, & Dieckmann, 2009).

Based on the evidence of these scales' predictive ability of actual decision-making, we found these scales as potential candidates to capture individuals' heterogeneity in responding to password nudges, which would make them promising measures for personalization of the nudges.

We performed two large-scale online experiments to evaluate the effects of personalized password nudges vs. using one-size-fits-all nudges. We first performed an exploratory study to examine how the effectiveness of the different password nudges correlates with the decision-making measures. We considered all these traits to be potential moderators, but did not formulate specific hypotheses about the exact relationships between the traits and the different nudges. Based on the relationships we did find, and specifically because we considered all of them as exploratory, we then performed another study in which we targeted nudges at the individuals we hypothesized would be most likely to benefit from each nudge. For this second study, we predicted that (a) decision-making style would consistently predict the effectiveness of the different nudges and, critically, (b) that personalizing which nudge is given to which individual would result in increased effectiveness of the nudges. Data files for both studies can be found at https://osf.io/n4h36.

Fig. 1 visually illustrates the workflow of developing and testing nudge personalization, as was applied in the current studies: Study 1 included an experimental test of the nudges and a measurement of the traits, enabling analyses of the nudge by trait interactions. Study 2 again measured the traits (on a new sample), and used the estimations of predicted effects of the nudges (from Study 1) on individuals with different profiles of traits, to personalize the assignment nudges and then to experimentally apply and test the effectiveness of the personalization of the nudges on passwords' strength.

2. Method

Participants. We recruited 2074 participants (52% males, M_{age} = 36.7, SD = 11.5) from Amazon's Mechanical Turk platform. Participants could take part if they lived in the U.S., had an approval rate of at least 95% on previous tasks (as recommended by Peer, Vosgerau, & Acquisti, 2014) and had completed an initial questionnaire, which involved completing the GDMS and NFC Scales. Due to the exploratory nature of this study, we recruited such a large sample in order to make certain enough power is obtained to detect the interactions between the nudges and the decision-making scales. A sensitivity power analysis, using G*Power software, showed that for an 80% power and alpha = 0.05 (two-tailed) this sample could detect an effect size as small as $f = 0.1$.

Design and procedure. Participants role-played a scenario in which they were asked to change the password for an email account. We told participants that the new password they create will also be used a week later to access the second stage of the study, for which they will receive a bonus payment. Passwords created during similar role-playing studies have been shown to be representative of passwords users create in the real world (Fahl, Harbach, Acar, & Smith, 2013; Komanduri et al., 2011). Nonetheless, because people often re-use existing passwords when asked to create them for studies (Egelman et al., 2013), all participants received a message stating that their chosen password was too weak, and that a new one would need to be selected, accompanied by one of the nudges. Control participants also received this message. Fig. 2 shows illustrations of the Control condition and the different password nudges. We did not enforce any minimum requirements (e.g., mandating characters or a minimum length). After participants created the new password, they filled out the Consideration for Future Consequences (CFC) scale (Strathman et al., 1994).

One week later, we sent invitations to participants to take part in a follow-up study. Out of the 2074 participants, 147 (7%) did not complete the second stage of the study. Additionally, we disqualified 103 (5%) responses that had duplicate IP addresses, retaining a final sample of 1824 participants. We gave participants three attempts to enter the passwords they had previously created, after which they were automatically allowed to proceed, regardless of what they entered. Then, participants filled out the Numeracy scale (Peters et al., 2006). We paid participants US$0.50 for completing the first stage (creating a password) and US$2 for completing the second stage (returning a week later).

The internal validity (Cronbach's α) of the decision-making measures was high for all scales: NFC = 0.944; GDMS: Intuitive = 0.846, Dependent = 0.837, Rational = 0.763, Avoidant = 0.915, Spontaneous = 0.843; CFC = 0.894; Numeracy = 0.615 (after removing item #7 due to

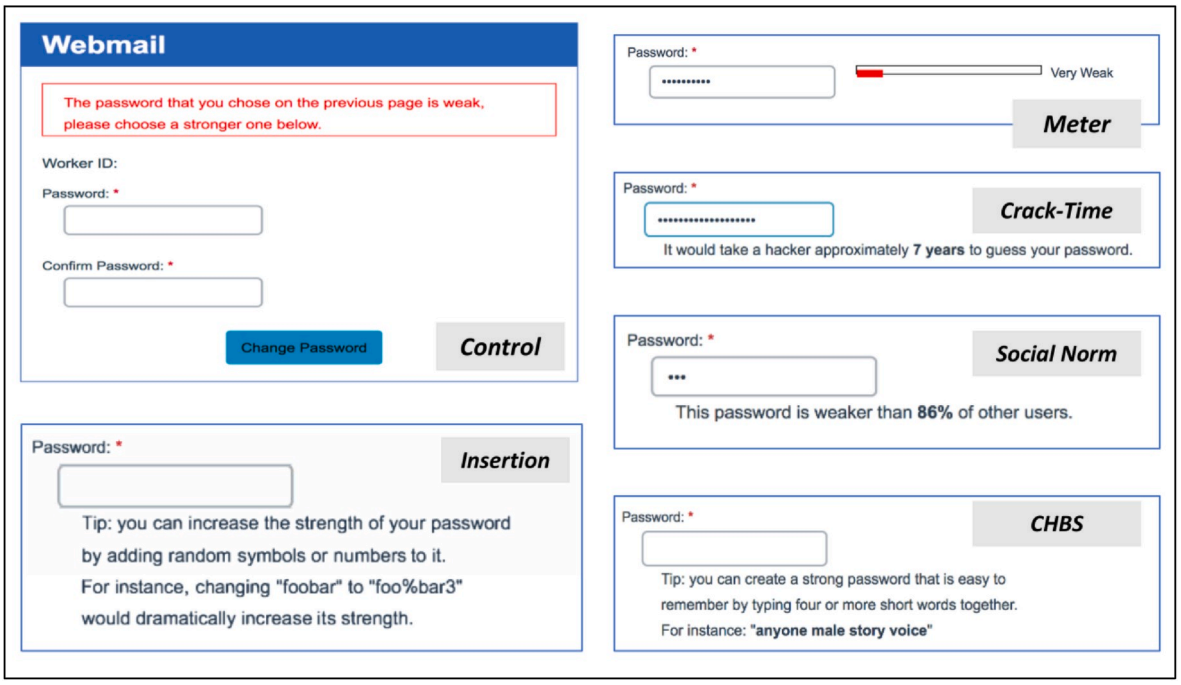


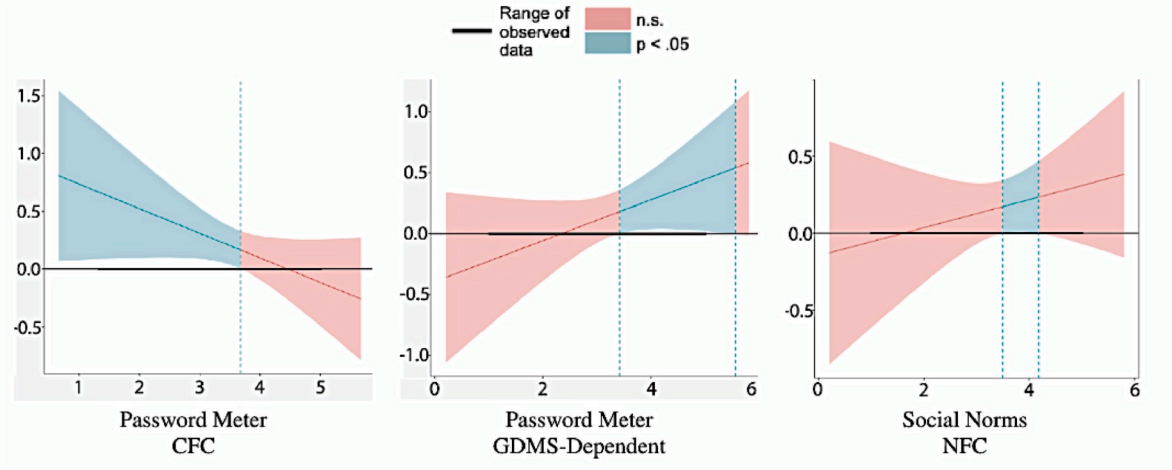
Fig. 2. Password creation page design and the password nudges used in the studies.

low variance).

3. Results

We examined the strength of the passwords participants ($N = 1842$) created after being asked to strengthen their passwords (i.e., post-nudge), comparing the four nudges and the control. We quantified password strength by an algorithm that uses a neural network to estimate the number of guesses it would take to crack a given password (Melicher et al., 2016). Given the extremely high variance of this metric, we log-transformed guess numbers. Our primary focus was on the interactions between traits and nudges, and we thus did not compare the overall effectiveness of the different nudges and first analyzed only the interactions of the nudges with the traits. Using the five scales of the decision-making traits, we analyzed each trait by nudge interaction using the Johnson-Neyman technique (Preacher, Curran, & Bauer, 2006). This technique uses bootstrapping methods to identify the "regions of significance"—ranges of trait values where the effect of each nudge on password strength is statistically significant. We applied this technique to each combination of nudge and trait separately and compared them to the control condition. This allowed us to examine the moderation effects of each trait on each nudge's effectiveness.

Our analyses uncovered significant moderation effects of traits on nudges. Fig. 3 shows how, for example, the password meter nudge was positively effective only for participants who had a relatively low CFC score or a relatively high score on the GDMS–Dependent Scale. As



Note: The cyan-colored areas show the region of trait values in which the nudge effect was statistically significant ($p < .05$), with a 95% confidence interval range around the slope of the nudge in the values of the trait.

Fig. 3. Moderating effects of decision-making traits on password nudges.

Table 1
Regions of significant moderation for each nudge in each trait. Numbers in the columns indicate the thresholds on the scale of each trait above, below or between which the effect of each nudge on password strength was statistically significant ($p < 0.05$). Values are confined to the possible range of the scale.

	Meter	Crack Time	Social Norms	CHBS
Numeracy		>8.59		>6.29
CFC	<3.68	2.11, 4.1		>1.92
NFC		3.02, 5.2	3.5, 4.18	1.41, 5.71
GDMS:				
-Intuitive		2.79, 4.63		<5.14
-Dependent	>3.41	2.79, 5.7		>2.1
-Rational		3.63, 4.85		>2.68
-Avoidant		1.96, 5		<5.64
-Spontaneous		>2.11		<4.35

another example, the effect of the social norm nudge was limited to those in the middle of the NFC Scale.

Table 1 details the values for the regions of significant moderation of the effect of the nudges, with a 95% confidence interval. For CHBS and Crack-Time, all traits showed significant moderation; for the Meter nudge, the CFC and GDMS–Dependent Scales showed significant moderation; for the Social nudge, NFC showed significant moderation; none of the traits was found to moderate the effects of the Insertion nudge. Figs. 4 and 5 show significant moderation effects for the Crack-Time and CHBS nudges, within a 95% confidence interval, respectively.

Note: The cyan-colored areas show the region of trait values in which the nudge effect was statistically significant ($p < 0.05$), with a 95% confidence interval range around the slope of the nudge in the values of the trait.

Although our focus was on the post-nudge password user created, we also examined their initial passwords. Looking at the overall sample, we found no statistically significant differences between the strength of the initial pre-nudge passwords between the conditions (Kruskal-Wallis $\chi^2 = 5.509$, $p = 0.357$, for both the raw or logged-standardized scores). As expected, we did find differences in the post-nudge passwords' strength between conditions (Kruskal-Wallis $\chi^2 = 47.727$, $p < 0.001$), as well as the change in strength pre- and post-nudge (Wilcoxon Signed Ranks test $p < 0.001$) for each condition. What this means is that passwords increased in strength after a nudge (including in the Control condition), though some nudges resulted in significantly greater effects than others.

We also examined individual differences in the strength of participants' initial passwords to determine whether users with high/low trait values provide stronger or weaker passwords than their counterparts. A multiple regression analysis on the standardized logged score of initial password strength, revealed significant differences ($F(3,1129) = 5.521$, $p < 0.001$, $R^2 = 0.012$) based on the measured traits. The strongest predictor of initial password strength was the GDMS-intuitive score that showed a negative correlation ($\beta = -0.09$, $p = 0.01$). NFC ($\beta = 0.07$, $p = 0.03$) and numeracy ($\beta = 0.04$, $p = 0.04$) were positively correlated with password strength. We found no significant interactions between these predictors. To summarize, users with higher numeracy or need for cognition, or lower reliance on intuition, chose stronger initial passwords.

We also checked whether the strength of the passwords participants entered could indicate that they regarded the task seriously, as if it is a real password creation situation. To do so, we compared the strength of all post-nudge passwords in our study to the strength of passwords in several leaked databases, scoring the password strength using the zxcvbn library. Fig. 6 shows the cumulative frequency of passwords being cracked in our dataset against four leaked databases from Ashley Madison, YouPorn, PHP Bulletin Board, and RockYou. As can be seen in Fig. 6, the password strength of our dataset was somewhat stronger than those in the leaked databases, suggesting our participants indeed regarded the password creation task seriously and entered realistic passwords.

As part of our analysis, we also tested the rate of recall of passwords between conditions. Across the participants who returned in our follow-up task a week later (*Phase Two*), we examined how many could remember the passwords that they created for the first part of the study. Overall, across all 1824 participants, 61.47% could successfully remember their passwords a week later, which is consistent with prior work (e.g., Komanduri et al., 2011). We conducted a chi-square test to evaluate whether the recall rates varied across conditions. We found that the p-value was not significant at the 0.05 level ($\chi^2 = 3.47$, $p = 0.63$), indicating that participants' recall rates were, on average, consistent across conditions. To test whether individual differences in the measured traits moderated recall levels, we ran a binary logistic regression with recall as the dependent variable, and the nudge group and trait measures as the predictors. We found that certain traits showed a main effect on recall levels: higher numeracy led to higher recall ($\beta =$

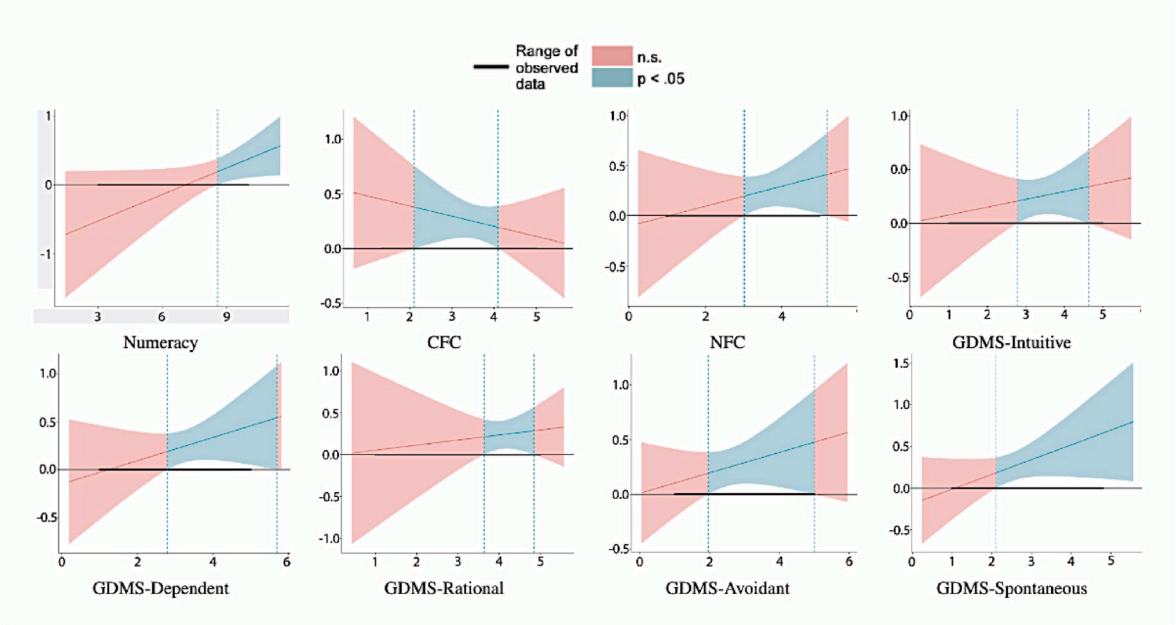


Fig. 4. Significant moderation effects of traits on the crack-time nudge.

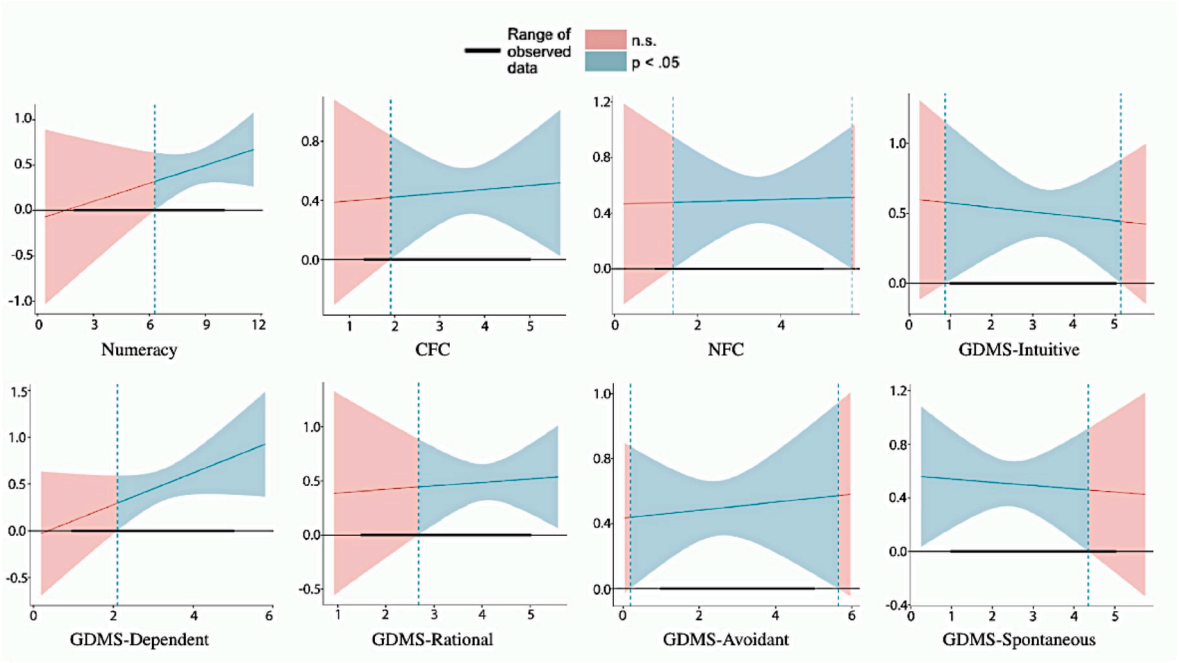


Fig. 5. Significant moderation effects of traits on the CHBS nudge.

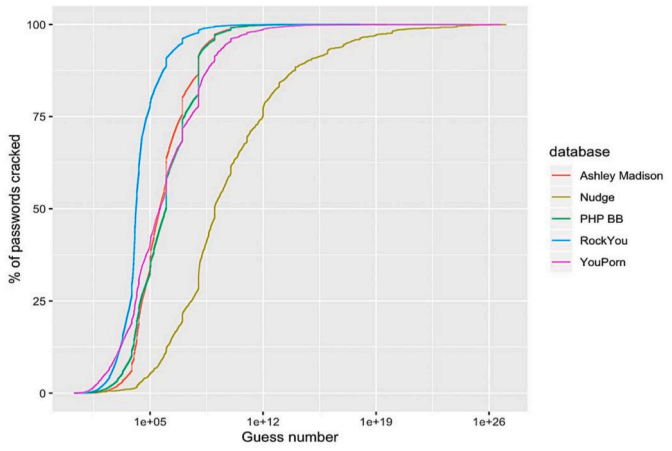


Fig. 6. Password strength in Study 1 sample (labeled "Nudge") vs. existing records of leaked databases.

0.095, SE = 0.042, p = 0.023); higher dependent GDMS led to lower recall (β = −0.21, SE = 0.079, p = 0.007); and higher spontaneous GDMS led to lower recall (β = −0.242, SE = 0.082, p = 0.003). However, we did not find any significant interaction between any of the traits and the nudges. This suggests that while some nudges had different effects on different users in terms of password strength (as presented earlier), this differential impact did not cause users to create more or less memorable passwords between nudges.

Overall, the results of this study suggest that decision-making style scales can be used to personalize and customize the nudges that would be presented to different individuals having different values for these traits. Specifically, it shows that if individuals' scores on one, some, or all of these traits are known, these scores could be used to predict which password nudge would result in the most optimal effect, i.e., the individual increasing the strength of his or her password. However, these findings only represent correlations, and not causal relationships and should be regarded with care. We designed our next study to validate the predictions made by the moderation analyses of decision-making styles on the nudges' effectiveness.

4. Method

Participants. We recruited 1146 participants from Mechanical Turk for the study. Participants had to have a greater than 95% approval rate, reside in the US, and did not participate in Study 1. We found that 215 did not respond correctly to an attention check question, leaving 931 valid responses. A sensitivity power analysis, using G*Power software, shows that for an 80% power and alpha = 0.05 (two-tailed) this sample could detect an effect size f as small as 0.1.

Design and procedure. We invited participants to complete the psychometric scales from Study 1 (GDMS, NFC, CFC, Numeracy). After these were completed, participants became eligible to participate in the main task, and were invited to do so on a later date. As in Study 1, we asked participants to role-play creating a password for an email service. Instructions suggested that the newly created password may be needed for a follow-up task; however, unlike in Study 1, we did not perform a follow-up task.

Using Monte-Carlo simulations based on the distributions of traits and effect sizes from Study 1, we computed for each participant the nudge that would be expected to produce the highest effect on the

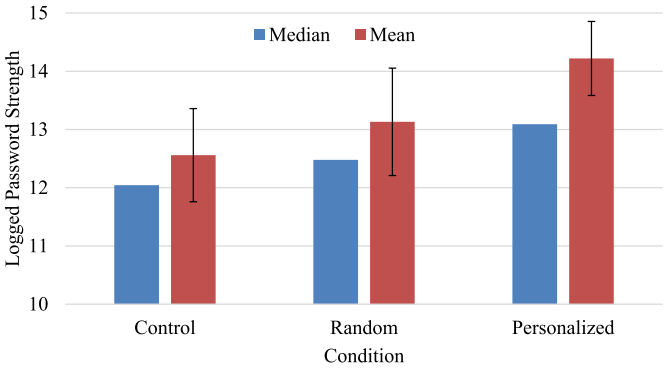


Fig. 7. Median and Mean Password Strength Between the Conditions (error bars show 95% confidence intervals around the means).

password strength of this participant, based on the participant's combined scores in the decision-making style scales. We focused on two of the more commonly used nudges: The Password Meter and the Crack-Time nudge. We allocated participants to three conditions. In the Control condition, participants were only told that their chosen password was weak and that they should choose a stronger one. In the Random condition, participants were randomly assigned to either the Crack-Time or Meter nudge, irrespective of their values on the decision-making scales. In the Personalized condition, participants received the nudge that was expected to produce the largest effect on their password strength, based on their scores on the decision-making scales. Our simulations estimated that the Crack-Time nudge would be optimal for 85% of the sample, whereas the Meter nudge would be optimal for 15% of the sample. Consequently, we oversampled participants into the Personalized condition so that a minimum number of participants would receive each nudge.

5. Results

We found that the mean password strength was highest for participants in the Personalized condition, compared to both the Control and Random conditions, $F(2, 920) = 5.201$, $p = 0.006$ (see Fig. 7). A planned contrast test of the Personalized vs. Random and Control conditions showed that the mean increase was statistically significant: $p < 0.001$. Additional analyses showed that the effect was mostly concentrated in the Crack-Time nudge, in which the difference in password strength between the Personalized and Random conditions was statistically significant, $p = 0.009$, whereas it was not statistically significant with the Meter nudge (see Fig. 8). The effect of the personalized nudge was also tested against the group of participants who were in the Random condition but only received the better of the two nudges – the Crack-Time nudge. This allowed us to compare how a personalized nudge would perform against the best-known nudge in this situation. We again found that a personalized nudge led to stronger passwords compared to those who received the "best" nudge, log standardized $M = 14.22$ vs. 12.61 (SD = 6.93, 6.74), t (586) = 2.34, $p = 0.02$, 95% $CI_{diff} = 0.26, 2.96$. Medians were also statistically different between these two groups (13.09 vs. 12.06), Mann-Whitney U test $p = 0.02$.

To quantify the effect size of personalizing the nudge, we employed the perspective of a realistic computer attacker and calculated the median amount of time and money it would take an attacker to crack passwords under the different conditions, based on the guess numbers estimated by the neural network (see details in the Appendix). As Table 2 shows, the median attack time was 4.2 times longer in the Personalized condition compared to the Random condition. Importantly, the difference between the Personalized and Random condition

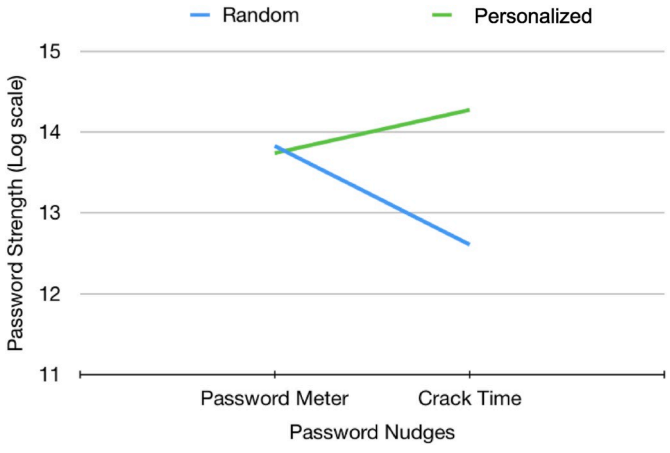


Fig. 8. Mean Password Strength for Crack-Time and Meter Nudges, When Given Randomly vs. Optimally to Participants.

Table 2
Attack time and cost estimates for average passwords generated under each condition (see Appendix for calculation details).

Condition	Attack time		Attack cost (US$)		
	Median/10 instances	Median/100 instances	Median	Mean	95% Upper CI
Control	6.66	0.67	$ 16,299	$ 1.98×10^{20}	$ 2.85×10^{20}
Random	18.49	1.85	$ 49,260	$ 2.17×10^{22}	$ 3.48×10^{22}
Personalized	74.27	7.43	$ 181,804	$ 1.96×10^{22}	$ 2.38×10^{22}

was larger than the difference between the Random and Control conditions: the attack time is only 2.8 times longer between these two conditions, and the attack cost is only 1.2 times larger compared to the Control. This means that, in this case, the relative marginal impact gained by using a *personalized* nudge was almost two times larger compared to the relative impact gained just by using *a* nudge.

6. General discussion

Personalizing nudges based on people's decision-making style resulted in passwords that, on average, quadrupled the resources (of either time or money) needed for an attacker to hack the password, compared to when personalization was not applied (or ten times harder to crack compared to when no nudge was given). While previous research either tested individual differences in nudge effectiveness post hoc (e.g., Costa & Kahn, 2013) or was satisfied in measuring the correlations between different types of messages to individual differences (e.g., Hirsh et al., 2012), our study is the first, to the best of our knowledge, to systematically test and compare the effectiveness of personalizing nudges, in the special yet ubiquitous domain of online password nudges.

Our findings show that personalizing nudges is not a small endeavor and requires certain steps for its successful implementation. First, one has to collect several nudges that have already been shown to be effective. In parallel, one has to identify individual traits and scales that could interact significantly with the effectiveness of these nudges. This we did in Study 1, which showed how correlations between existing password nudges and individual differences in decision-making styles could be used in order to identify profiles based on the specific relationships found. Using these profiles, one then needs to run simulations that could provide predictions on which nudges to administer to the different profiles identified. This we did in Study 2, where we showed that when such relationships are well-defined and used to determine, ex ante, which individual receives which nudge, the result is a substantial improvement in the effectiveness of the nudge, compared to when the same nudge is given in a one-size-fits-all approach. Because the relationships found in the first steps (Study 1) might be spurious or coincidental, the latter steps of experimentally testing the marginal benefits of personalization (Study 2) are critical, and we thus regard both studies as necessary steps for the such of the personalized nudges approach.

These findings have clear implications for both researchers and practitioners of nudges, within and beyond the fields of cybersecurity. Using our approach and the methodology we outline and test in our studies, the design, evaluation and implementation of nudges can be considerably improved to maximize effectiveness. This approach seems most feasible in the domain of online nudges, where individuals' details could be collected and used to customize and personalize the nudges they receive. In such domains, nudge personalization can reach far beyond online security and password strength into important areas of health, savings, environmental protection, and more. For example, in the area of savings, a large field study showed that using a default

allocation in the online tax form, US tax-payers could be nudged to pledge some of their annual tax return for savings (Grinstein-Weiss, Russell, Gale, Key, & Ariely, 2017). Given the large amount of relevant financial information provided by the applicant on that same tax form, it is highly likely that the default allocation amount could have been personalized to fit each individual, encouraging them to save more, while still keeping enough for spending.

Important caveats, however, must be considered before applying nudge personalization. First, we do not mean to advocate the use of personalized nudges in any situation. Specifically, for the area of computer security, we are hesitant to recommend their use every time users are asked to create passwords. That is because if every website would nudge users to construct the strongest possible passwords, users might be less likely to remember all of their passwords from the different web sites. Furthermore, current best practices suggest that users should not be choosing passwords at all, but instead using password managers to automatically generate random unique passwords for each service. Second, as the news about Facebook and Cambridge Analytica has shown us, individuals' personal, and sometimes sensitive, data are already being used for personalization of less benevolent interventions. For nudge personalization, users' information (e.g., traits) must be available or users' trait scores implicitly inferred from observations of their behaviors (Kosinski, Stillwell, & Graepel, 2013; Matz, Kosinski, Nave, & Stillwell, 2017). To preserve privacy, such information must be stored in trusted locations on users' devices or browsers, so that the information about users' traits or scores will never be revealed to intermediaries or third parties, and can be used effectively and securely for nudge personalization. Nevertheless, in situations where personal information is already being collected and stored (such as the examples mentioned above), this information should be used in a manner that could increase consumers' welfare, thereby somewhat mitigating the harm to privacy caused by collecting and storing consumers' personal information.

Even more broadly, nudge personalization raises important ethical considerations that must be taken into account by policy makers. It is not clear, ex-ante, when and how should nudges be personalized to achieve different outcomes and further research and theoretical development is necessary in order to establish a framework or guidelines on when, how and to what extent should nudges be personalized or not. Like any other persuasion method, personalized nudges might also be used in sinister motives by less-benevolent or less-democratic governments or organizations (for example, to reduce instead of increase voting rates). People with low income, for example, could be particularly vulnerable to nudges personalized to take advantage of their cognitive scarcity which has been found to hamper their optimal decision-making abilities (fir_2013, Mullainathan_and_Shafir_2013, Mullainathan & Shafir, 2013). As should be the case with almost any public policy, nudge personalization must also undergo scrutinized evaluations to show that it indeed leads to better outcomes and improved welfare compared to the alternatives of deploying only one "average" nudge, or not nudging at all.

Our research suggests that personalization of nudges holds out promise of being able to better the welfare of individuals in several respects. First of all, as already demonstrated, it can increase the overall effectiveness of the nudge. But more than that, it may also help mitigate the risks of harming the welfare of some sub-groups in the population. In fact, even when a nudge shows an average positive effect in the overall sample, this average may comprise of many combinations of effect sizes different in magnitude and direction in different sub-groups: some groups could be less affected by the nudge, some groups could be indifferent to the nudge, and some groups could even be negatively (and even strongly so) affected by the nudge. In the latter case, it is obvious that using the nudge could be objectionable from a managerial and policy perspective as causing unintended consequences on, for example, a minority or under-protected group in the organization or the society. Research to date has generally neglected to consider such effects of heterogeneity, but that does not mean they do not occur or could not manifest in the future. Our approach of personalizing nudges according to people's individual traits ensures not only that the different individuals are nudged in the way most appropriate for them, but that other individuals would be kept safe from harmful effects of a certain nudge, if and when such harms might occur.

To conclude, we argue that, based on our findings, managers and policy makers should shift their focus away from homogeneously deploying nudges in one-size-fits all approaches. Instead, efforts should be directed at developing policies that would ensure different individuals receive different nudges in a way that is most effective for each one of them. This is no small feat, and solving the above challenges requires a fair amount of coordination among researchers, practitioners, managers and policy makers. However, we are confident that this goal is a critical step forward in the development and application of nudges and behavioral public policy to improve people's decisions and welfare.

CRediT authorship contribution statement

Eyal Peer: Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Serge Egelman:** Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Marian Harbach:** Conceptualization, Methodology, Formal analysis, Investigation, Writing - review & editing. **Nathan Malkin:** Conceptualization, Methodology, Formal analysis, Investigation, Writing - review & editing. **Arunesh Mathur:** Methodology, Formal analysis, Investigation, Writing - review & editing. **Alisa Frik:** Investigation, Writing - review & editing.

Acknowledgments

This study was sponsored by NSF Award #1528070 and by BSF Award # 2014626. We thank the people at the International Computer Science Institute at Berkeley, CA, and Tamar Ben-Meir, for their assistance with this research project.

Appendix. Attack time and cost calculations in Study 2

To estimate the resistance of passwords against a realistic attacker, we estimate the time and cost of guessing the median password within each condition. Given that the one-way cryptographic hash-function that is used to protect user passwords at rest on a server is not broken in itself, the best strategy for the attacker is to guess what the password could be, apply the same hash function and compare the output to the password hash obtained in a data breach. How many guesses it will take an attacker to crack a given password is estimated by the neural network password strength estimator described in the paper. We use these guess numbers for each condition as the basis to compare the strength of the passwords obtained under our experimental modifications. The median number of guesses for the final passwords per condition in our sample were as follows: Control = 1.11 × 10^12; Random = 3.07 × 10^12; Optimal = 12.34 × 10^12.

To simulate a realistic attacker, we assume access to a number of the largest (p3.16xlarge) Amazon Web Services GPU instances, which cost $10.2 each per hour with a long-term contract (see: https://aws.amazon.com/ec2/instance-types). Benchmarks estimate that version 4.0.0 of hashcat can calculate about 192,300 password hashes with 100,000 rounds of PBKDF2-HMAC-SHA256 on one of those instances. Given the number of guesses for each password under the conditions and guessing ability per server instance, Table 2 shows the amount of time and money the attacker would have to invest to crack the median password. Given the nature of Amazon Web Services, the attacker could horizontally scale the attack using approximately the same amount of money. We, therefore, argue that attack cost is currently the main limiting factor for dedicated attackers. We provide attack time as a reference for attackers that have free access to limited server resources and are thus limited in time.

Submitted Manuscript: Confidential.

References

Acquisti, A., Adjerid, I., Balebako, R., Brandimarte, L., Cranor, L. F., Komanduri, S., & Wang, Y. (2017). Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys, 50*(3), 44. https://doi.org/10.1145/3054926.

Akhawe, D., & Felt, A. P. (2013). Alice in warningland: A large-scale field study of browser security warning effectiveness. In *Presented as part of the 22nd {USENIX} security symposium ({USENIX} security 13)* (pp. 257–272).

Appelt, K. C., Milch, K. F., Handgraaf, M. J., & Weber, E. U. (2011). The decision making individual differences inventory and guidelines for the study of individual differences in judgment and decision-making research. *Judgment and Decision Making, 6*(3), 252–262.

Bazerman, M. H., & Gino, F. (2012). Behavioral ethics: Toward a deeper understanding of moral judgment and dishonesty. *Annual Review of Law and Social Science, 8*, 85–104.

Breman, A. (2011). Give more tomorrow: Two field experiments on altruism and intertemporal choice. *Journal of Public Economics, 95*(11–12), 1349–1357.

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*(3), 306–307.

Cacioppo, J. T., Petty, R. E., & Morris, K. J. (1983). Effects of need for cognition on message evaluation, recall, and persuasion. *Journal of Personality and Social Psychology, 45*(4), 805.

Cai, C. W. (2019). *Nudging the financial market? A review of the nudge theory*. Accounting & Finance. https://doi.org/10.1111/acfi.12471.

Camilleri, A. R., & Larrick, R. P. (2014). Metric and scale design as choice architecture tools. *Journal of Public Policy and Marketing, 33*(1), 108–125.

Carroll, G. D., Choi, J. J., Laibson, D., Madrian, B. C., & Metrick, A. (2009). Optimal defaults and active decisions. *Quarterly Journal of Economics, 124*(4), 1639–1674.

Chellappa, R. K., & Sin, R. G. (2005). Personalization versus privacy: An empirical examination of the online consumer's dilemma. *Information Technology and Management, 6*(2–3), 181–202.

Collier, C. A. (2018). July). Nudge theory in information systems research A comprehensive systematic review of the literature. In *Academy of management proceedings* (Vol. 2018, p. 18642). Briarcliff Manor, NY 10510: Academy of Management.

Costa, D. L., & Kahn, M. E. (2013). Energy conservation "nudges" and environmentalist ideology: Evidence from a randomized residential electricity field experiment. *Journal of the European Economic Association, 11*(3), 680–702.

Dalal, R. S., & Bonaccio, S. (2010). What types of advice do decision-makers prefer? *Organizational Behavior and Human Decision Processes, 112*(1), 11–23.

Damgaard, M. T., & Nielsen, H. S. (2018). Nudging in education. *Economics of Education Review, 64*, 313–342.

Egelman, S., & Peer, E. (2015). The myth of the average user: Improving privacy and security systems through individualization. In *In Proc. 2015 new security paradigms workshop*. The Netherlands: ACM, Twente. https://doi.org/10.1145/2841113.2841115. ACM.

Egelman, S., Sotirakopoulos, A., Muslukhov, I., Beznosov, K., & Herley, C. (2013). Does my password go up to eleven?: The impact of password meters on password selection. In *Proc. SIGCHI conf. Human factors in computing systems* (pp. 2379–2388). Paris, France): ACM. https://doi.org/10.1145/2470654.2481329.

Fahl, S., Harbach, M., Acar, Y., & Smith, M. (2013). On the ecological validity of a password study. In *Proceedings of the ninth symposium on useable privacy and security* (p. 13). ACM.

R, W., Felt, A. P., Consolvo, S., Malkin, N., Thompson, C., & Egelman, S. (2018). An experience sampling study of user reactions to browser warnings in the field. In *Proceeding of the 2018 CHI confernece human factors in computing systems* (p. 512). Montreal, QC, Canada): ACM.

Forget, A., Chiasson, S., van Oorschot, P. C., & Biddle, R. (2008). Improving text passwords through persuasion. In *Proceedings of the 4th symposium on useable privacy and security* (pp. 1–12). ACM.

Goldstein, D. G., Johnson, E. J., Herrmann, A., & Heitmann, M. (2008). Nudge your customers toward better choices. *Harvard Business Review, 86*(12), 99–105.

Grinstein Weiss, M., Russell, B. D., Gale, W. G., Key, C., & Ariely, D. (2017). Behavioral interventions to increase tax-time saving: Evidence from a national randomized trial. *Journal of Consumer Affairs, 51*(1), 3–26.

Halpern, D. (2016). *Inside the nudge unit: How small changes can make a big difference*. London, UK: Random House.

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

Hauser, J. R., Urban, G. L., Liberali, G., & Braun, M. (2009). Website morphing. *Marketing Science, 28*(2), 202–223.

Hirsh, J. B., Kang, S. K., & Bodenhausen, G. V. (2012). Personalized persuasion: Tailoring persuasive appeals to recipients' personality traits. *Psychological Science, 23*(6), 578–581.

Jeske, D., Coventry, L., Briggs, P., & van Moorsel, A. (2014). *Nudging whom how: IT proficiency, impulse control and secure behavior. Personalizing behavior change technologies workshop, toronto, Canada* Accessed February 2, 2019.

Johnson, E. J., & Goldstein, D (2003). Do defaults save lives? Science, 302(5649), 1338-1339.

Karlan, D., McConnell, M., Mullainathan, S., & Zinman, J. (2016). Getting to the top of mind: How reminders increase saving. *Management Science, 62*(12), 3393–3411.

Komanduri, S., Shay, R., Kelley, P. G., Mazurek, M. L., Bauer, L., Christin, N., & Egelman, S. (2011). Of passwords and people: Measuring the effect of password-composition policies. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2595–2604). ACM.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences, 110*(15), 5802–5805.

Malkin, N., Mathur, A., Harbach, M., & Egelman, S. (2017). *Personalized security messaging: Nudges for compliance with browser warnings*. Second *European Workshop on Usable Security*. Internet Society.

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences, 114*(48), 12714–12719.

Melicher, W., Ur, B., Segreti, S. M., Komanduri, S., Bauer, L., et al. (2016). Fast, lean, and accurate: Modeling password guessability using neural networks. In *USENIX security symposium* (pp. 175–191).

Milkman, K. L., Beshears, J., Choi, J. J., Laibson, D., & Madrian, B. C. (2011). Using implementation intentions prompts to enhance influenza vaccination rates. *Proceedings of the National Academy of Sciences, 108*(26), 10415–10420.

Mullainathan, S., & Shafir, E. (2013). *Scarcity: Why having too little means so much*. New York: Henry Holt and Company.

Organisation de Coopération, et al.Développement Economiques. (2017). *Behavioural Insights and public policy: Lessons from around the world*. OECD Publishing. Retrieved on 1/1/2019 from http://www.oecd.org/gov/regulatory-policy/behavioural-insights-and-public-policy-9789264270480-en.htm.

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods, 46*(4), 1023–1031.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science, 17*(5), 407–413.

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics, 31*(4), 437–448.

Quigley, M. (2013). Nudging for health: On public policy and designing choice architecture. *Medical Law Review, 21*(4), 588–621.

Reisch, L. A., & Sunstein, C. R. (2016). Do Europeans like nudges? *Judgment and Decision making, 11*(4), 310–325.

Renaud, K., & Zimmermann, V. (2019). Nudging folks towards stronger password choices: Providing certainty is the key. *Behavioural Public Policy, 3*(2), 228–258.

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135*(6), 943.

Scott, S. G., & Bruce, R. A. (1995). Decision-making style: The development and assessment of a new measure. *Educational and Psychological Measurement, 55*(5), 818–831.

Shay, R., Kelley, P. G., Komanduri, S., Mazurek, M. L., Ur, B., Vidas, T., & Cranor, L. F. (2012). July). Correct horse battery staple: Exploring the usability of system-assigned passphrases. In *Proceedings of the eighth symposium on useable privacy and security* (p. 7). ACM. https://doi.org/10.1145/2335356.2335366.

Shillair, R., Cotten, S. R., Tsai, H. Y. S., Alhabash, S., LaRose, R., & Rifon, N. J. (2015). Online safety begins with you and me: Convincing Internet users to protect themselves. *Computers in Human Behavior, 48*, 199–207.

Smith, S. M., & Levin, I. P. (1996). Need for cognition and choice framing effects. *Journal of Behavioral Decision Making, 9*(4), 283–290.

Strathman, A., Gleicher, F., Boninger, D. S., & Edwards, C. S. (1994). The consideration of future consequences: Weighing immediate and distant outcomes of behavior. *Journal of Personality and Social Psychology, 66*(4), 742.

Sunstein, C. R. (2013). Impersonal default rules vs. Active choices vs. Personalized default rules: A triptych. Available at: SSRN: https://ssrn.com/abstract=2171343.

Szaszi, B., Palinkas, A., Palfi, B., Szollosi, A., & Aczel, B. (2018). A systematic scoping review of the choice architecture movement: Toward understanding when and why nudges work. *Journal of Behavioral Decision Making, 31*(3), 355–366.

Thaler, R. H., & Benartzi, S. (2004). Save more tomorrow™: Using behavioral economics to increase employee saving. *Journal of Political Economy, 112*(S1), 164–S187.

Thaler, R. H., & Sunstein, C. (2008). *Nudge: Improving Decisions about health, wealth and happiness*. New Haven, CT: Yale University Press.

The Behavioral Insights Team. (2014). East: Four simple ways to apply behavioural Insights. Retrieved at November 1, 2019 from https://www.bi.team/wp-content/uploads/2015/07/BIT-Publication-EAST_FA_WEB.pdf.

Ur, B., Kelley, P. G., Komanduri, S., Lee, J., Maass, M., Mazurek, M. L., & Christin, N. (2012). How does your password measure up? The effect of strength meters on password creation. *USENIX security symposium* (pp. 65–80).

Vance, A., Eargle, D., Ouimet, K., & Straub, D. (2013). Enhancing password security through interactive fear appeals: A web-based field experiment. In *2013 46th Hawaii international conference on system sciences* (pp. 2988–2997). IEEE.

VanEpps, E. M., Downs, J. S., & Loewenstein, G. (2016). Calorie label formats: Using numeric and traffic light calorie labels to reduce lunch calories. *Journal of Public Policy and Marketing, 35*(1), 26–36.

Vesanen, J. (2007). What is personalization? A conceptual framework. *European Journal of Marketing, 41*(5/6), 409–418.

Wheeler, D. L. (2016). zxcvbn: Low-Budget password strength estimation. In *USENIX security symposium* (pp. 157–173). Retrieved on November 1, 2019 from https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_wheeler.pdf.

Zimbardo, P. G., Keough, K. A., & Boyd, J. N. (1997). Present time perspective as a predictor of risky driving. *Personality and Individual Differences, 23*(6), 1007–1023.