Lessons Learned from Evaluating Eight Password Nudges in the Wild

Karen Renaud
Abertay University
k.renaud@abertay.ac.uk

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

Joseph Maguire & Steve Draper
University of Glasgow
{joseph.maguire,steve.draper}@glasgow.ac.uk

Abstract

Background. The tension between security and convenience, when creating passwords, is well established. It is a tension that often leads users to create poor passwords. For security designers, three mitigation strategies exist: issuing passwords, mandating minimum strength levels or encouraging better passwords. The first strategy prompts recording, the second reuse, but the third merits further investigation. It seemed promising to explore whether users could be subtly *nudged* towards stronger passwords.

Aim. The aim of the study was to investigate the influence of visual nudges on self-chosen password length and/or strength.

Method. A university application, enabling students to check course dates and review grades, was used to support two consecutive empirical studies over the course of two academic years. In total, 497 and 776 participants, respectively, were randomly assigned either to a control or an experimental group. Whereas the control group received no intervention, the experimental groups were presented with different visual nudges on the registration page of the web application whenever passwords were created. The experimental groups' password strengths and lengths were then compared that of the control group.

Results. No impact of the visual nudges could be detected, neither in terms of password strength nor length. The ordinal score metric used to calculate password strength led to a decrease in variance and test power, so that the inability to detect an effect size does not definitively indicate that such an effect does not exist.

Conclusion. We cannot conclude that the nudges had no effect on password strength. It might well be that an actual effect was not detected due to the experimental design choices. Another possible explanation for our result is that password choice is influenced by the user's task, cognitive budget, goals and pre-existing routines. A simple visual nudge might not have the power to overcome these forces. Our lessons learned therefore recommend the use of a richer password strength quantification measure, and the acknowledgement of the user's context, in future studies.

1 Introduction

The first encounter with a new system or service, for many individuals, requires the creation of a password. This authentication approach is based on the possession of some secret shared knowledge, known only to the user and this one system.

People are asked to provide passwords so frequently, and inconveniently, that they end up choosing weak passwords, leaving themselves vulnerable to attack [30]. In effect, password choice becomes something of an obstacle to be hurdled in order to be able to satisfy legitimate goals. The primary problem is the fact that memory limitations tug people towards memorable and predictable secrets, whereas strong security mandates more effort. Strength can be achieved either by using a hard-to-remember and hard-to-guess nonsense string, or by using a long pass phrase. Both are personally more costly than a weak password.

Some believe that we should simply enforce strong passwords [15] or expire passwords regularly [18]. The problem is that neither the former nor the latter guarantee increase resistance to attack [53, 57]. Moreover, restrictive, complex password policies aimed at mandating strong passwords can conflict with users' needs, increase effort and ultimately compromise productivity and security [24, 48, 52].

The other option is to replace the password with something like a biometric or token-based authentication [5, 38]. Neither of these is perfect either. No biometric is ubiquitous and infallible [32] and tokens are expensive and easily lost or stolen.

Other alternatives are graphical passwords, mnemonic passwords or passphrases [1, 51, 28] but these have not really gained widespread acceptance and even passphrases have their flaws [29, 39].

While many focus on the password's deficiencies, it must be acknowledged that passwords also have advantages. They are easy to deploy, accessible to those with disabilities, cost-effective, preserve privacy and are easily replaced [6].

Instead of focusing myopically on the password choice event, we should contemplate password creation as one component of an entire authentication eco-system, and consider that the end user needs more support throughout the process. Horcher and Tejay [23] claim that users are poorly scaffolded during the password creation process, and that this contributes to poor password choice. Solutions that scaffold by offering dynamic feedback on password quality are designed to encourage deliberation and reflection during password creation [17]. However, such approaches have, thus far, not significantly improved the quality of passwords [12, 17, 45].

There is increasing evidence that behaviour can be influenced through surprisingly small and inexpensive interventions called "nudges" [21].

Transferring successful nudges from other areas to the authentication context was something we wanted to test to find out whether these would encourage users to create stronger passwords. The hypothesis we tested was:

> **H1:** The presence of a visual nudge will lead to longer and stronger passwords.

We carried out a longitudinal study to investigate the potential of eight different user interface nudges, displayed during password creation, calculated to influence password choice. The contributions of this paper are:

- Details of how nudges were tested in the wild, and the ethical constraints we encountered.
- Empirical evidence that the tested nudge conditions did not significantly impact password quality.
- Reflection on the results and suggested explanations for the negative finding reported by the study.

The paper concludes with a discussion of lessons learned and recommendations for future studies of this kind.

2 Background

A nudge can be considered a mechanism that guides individuals to make wiser choices without their necessarily being aware of its influence [34]. An intervention can only be considered a nudge if individuals are able easily

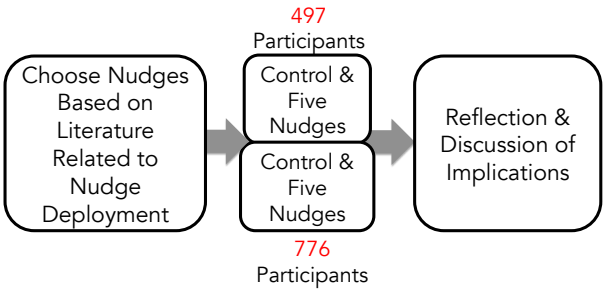

Figure 1: Results Reported in this Paper

to resist its influence [43]. A good example of a nudge is the house-fly painted on urinals in an Amsterdam airport. This nudge had the desired effect of reducing spillage, but could equally have been ignored by urinal users.

The subtle nudge approach has proved popular with western governments [44, 22], who have adopted nudges in key areas such as tax and public health [50]. A small alteration in letter text sent to individuals significantly improved tax payment rates [21]. However, such use has been criticized with the suggestion that nudges do not promote long-term behaviour change [36]. Nevertheless, this may not be an issue for use in authentication if the motivation is to promote optimal decisions at the moment of password creation.

There is an argument that people sometimes create passwords unthinkingly, basically operating using their *autopilot* (System 1) thinking, rather than deliberately engaging (System 2) level thinking to choose a good password [43]. Sunstein [41] explains that nudges can work in tandem with educational efforts by impacting System 1 thinking, with educational efforts targeting System 2 thinking, thus complementing each other.

Jeske *et al.*. [26] demonstrated such an approach when it came to nudging users to select the most secure wireless networks. They found that nudges could be effective, but that personal differences also played a role in the security decisions. Similarly, Yevseyeva *et al.* [56] experimented with nudging people towards secure wireless network selection using different variations of a prototype application. The found a combination of colour coding and the order in which the Wi-Fi networks were listed to be most effective.

Nudges have also been deployed to improve decisions surrounding privacy. Choe *et al.* [10] investigated positive and negative framing of privacy ratings to nudge individuals away from privacy-invading applications. They demonstrated that framing, as well as user ratings, had the potential to nudge individuals towards privacy-respecting applications. Similarly, Balebako *et al.* [3] suggest that nudges can support users in making

more optimal decisions in privacy when it comes to location sharing. They argue that individuals, left unaided, might well make regrettable privacy decisions due to the cognitive load caused by having to consider all possible ramifications of a single privacy decision. Similarly, Almuhimedi *et al.* [2] investigated user awareness of privacy invasion by making usually invisible data sharing, visible . Almuhimedi *et al.* demonstrated that the majority were nudged to reassess their privacy permissions when data was presented.

Authentication nudge studies have delivered disappointing results so far [12, 17]. One study attempted to exploit the Decoy Effect [27]. This design involves giving users three choices: one inferior, one very expensive, and a middle-of-the-road option that designers want people to choose. The decoy study [45] offered users their own password choice, a complex hard-to-remember password and the alternative they really wanted users to choose: a long and memorable password. The relative strengths of the three passwords was displayed to influence choice. The results were disappointing [45].

Another nudge effort that has enjoyed much research attention is the password strength meter. These mechanisms provide strength feedback, either post-entry or dynamically. Mechanisms can provide colour indicators, strength indicator bars, or informative text [8].

Ur *et al.* [47] compared a number of different password strength meters and discovered that meters impacted password strength. However, they tested their meters using a Mechanical Turk survey. The fact that the created passwords carried no cost might have led to respondents formulating somewhat unrealistic passwords. Ur *et al.*'s study was an essential first step in exploring these kinds of interventions, giving us hope that nudges could be designed to work in the wild too.

Sotirakopoulos [40] attempted to influence password choice by providing dynamic feedback. No difference between a horizontal strength meter and the comparison to peer passwords emerged. Vance *et al.* [49] also reported that password strength meters only impacted password strength in conjunction with an interactive fear appeal treatment condition that included a message on the seriousness of the threat. An interactive password strength meter and a static fear appeal did not impact password strength.

Egelman *et al.* [17] did test the impact of providing password meters in the wild. They found that the meters made no observable difference to password choice, unless users perceived the account to be important. If people *do not* attribute value, then it is understandable that the password meter makes no difference to their choice.

Privacy nudges have been more successful than authentication nudges so far. Privacy choices, however, entail people having a choice between two fairly equivalent options [10, 26]. Nudging in authentication does not match this pattern of use and, in fact, initial studies on nudges in authentication have delivered mixed results, as described above. Still, nudges have been successfully deployed in other application areas, and at least two explanations for the lack of success in authentication. It might be the case that authentication is unsuited to nudging influence. On the other hand, it could be that a success authentication nudge is yet to be discovered discovered.

Much nudging in authentication has focused on password strength meters. We thus carried out a study to extend the evidence base by testing a number of visual authentication nudges. We tested nudges which focus on cognitive effects (e.g. social norms and expectation) that have rarely been tested in the authentication context.

We displayed different visual nudges during password creation events, in order to determine whether they exerted any influence over users during password creation.

3 Method

Current efforts to improve password choice focus primarily on the individual. However, situational and contextual influences could minimise the impact of individually-focused interventions [31]. Furthermore, social influence is a strong driver of compliance [11, 35]. Interventions could conceivably exploit the power of social norms to influence individual behaviours [4]. Since our target users in this study were students this context includes the University and their School. Visual nudge figures were created beforehand and displayed statically to ensure that all students saw exactly the same image. A dynamically-updated image might have confounded results because participants would then have seen different images, confounding our results. We designed one nudge for each cognitive effect we tested and that has led to positive results in other research areas.

Due to the exploratory and "in the wild" nature of this study, we decided to evaluate a range of cognitive effects with one nudge each, instead of focusing on one effect and creating several variations of nudges to exert influence in that one area. If a positive impact resulted, further exploration of the effect and variations of the nudge would be a direction for future research.

3.1 The Nudges

We conducted two studies with a similar experimental design: In each of the two studies the nudges served as an independent variable with six levels, a control group that did not receive any intervention and five different nudge conditions. From the nudges described below, nudges N1 to N5 were tested in study 1. Nudges N6 to N8 were tested in study 2 along with a replication of N2 and

N3. The dependent variables were (1) password strength measured with the strength estimator `zxcvbn.js` [54] and (2) password length (for further information see the Apparatus section). All nudges were presented to the participants on the registration page of a web-based university application that is described in the Apparatus Section below.

- **N0: Control**. The control group was presented with the standard registration page which asked users to "Choose a Password".
- **N1: Subconscious Mind.** *Testing the Priming Effect*. In authentication people are almost always prompted to provide a new password with the words: "Choose a Password". It is possible that this phrase could be partly responsible for one of the most common passwords being "password". If this admittedly subtle prime is a causative we ought to be able to influence choice by changing the word to "secret", and then see how many participants choose the password 'secret'. Thus, in our first experimental condition the phrase "*Choose a Password*" was replaced with "*Choose a Secret*".
- **N2: University Context**. *Testing the Expectation Effect*. Instead of mandating password strength requirements, the participants were shown the static graphic (Figure 2) that suggests that their password ought to be stronger than the average password chosen by other students.

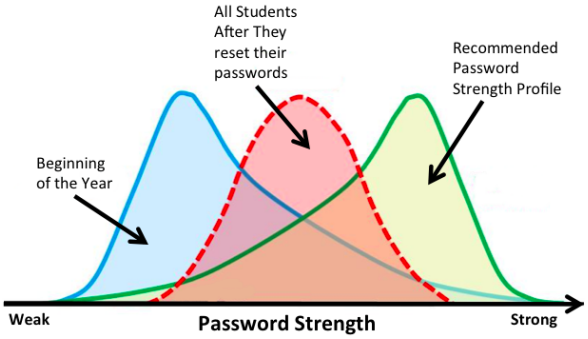

Figure 2: Expectation Effect Nudge Graph [37]

- **N3: School Context**. *Testing the Strength of In-Group Effect*. We suggested that participants identify themselves with students within their school, referred to as SoCS (Figure 3) in the graphic that was shown to them.

Some people argue that people do not know how strong their passwords are. To determine whether dynamic feedback reflecting the strength of their passwords would make a difference, we superimposed the arrow shown in Figure 4 over the images in Figures 2 and 3, giving us conditions N4 and N5. The strength feedback was based on the same strength estimator `zxcvbn.js` [54] that was used to calculate password strength in all experimental conditions (see section Apparatus for further details).

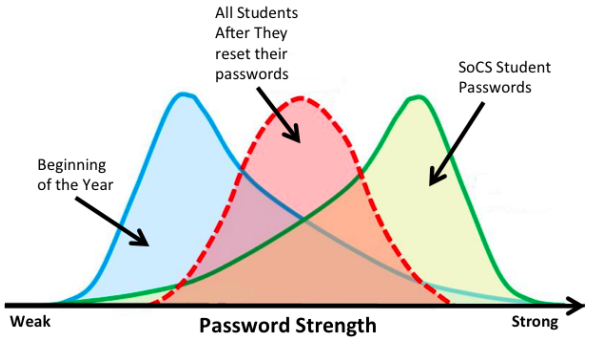

Figure 3: The In-Group Nudge Graph [9]

Figure 4: Strength Indicator

- **N4: University Context & Feedback**. Testing the combination of the Expectation Effect graph, with an interactive password strength meter superimposed over it. This would theoretically allow the user to see where on the x-axis their password is located, in terms of strength, as they entered it.
- **N5: School Context & Feedback**. Testing the combination of the In-Group Effect graph, with same dynamic strength feedback indicator as N4.
- **N6: Social Norm**. An image of eyes on a wall, appearing to "watch", makes people more likely to pay into an honesty box and also has the potential to reduce littering [4]. Given the impact of displayed eyes in other fields we considered it worthwhile to test whether the perception of being watched would encourage stronger passwords we displayed a pair of eyes above the password entry field.

For the final two conditions we asked the participants to reflect on the strength of their passwords to make them

pause and think about the password. Due to the constraints imposed by the ethics committee, no self-report free-form text was available. Instead, the participants were asked to rate the perceived strength of their password on a scale below the Figures as shown in 2 and 3 respectively, giving us conditions seven and eight.

This was intended to drive processing up to the System 2, deliberate level, of processing, to offset the automaticity they might be subject to while choosing passwords.

- **N7: University Environment & Reflection**. This treatment displayed the same image as N2, and asked the user to rate the strength of the password he or she had just entered. The instruction referred to them as 'a student' in order to highlight their University affiliation;
- **N8: School Environment & Reflection**. This group displayed the same image as N3, in addition to asking the user to rate the strength of their password. The instruction referred to them as 'a computing science student', once again to emphasise their in-group affiliation.

Apparatus. The nudges were tested using a web-based university application where students were provided with coursework deadlines, timetable information and project allocations. The authentication scheme was based on standard alphanumeric authentication, i.e. a username and a password.

We did not enforce a password policy nor a time limit for password creation as we wanted to test the sole impact of the nudges on password creation. However, the university where the study took place generally suggests that passwords should be at least eight characters long (passphrases are recommended), include at least one non-letter and should be changed at least once a year). Access to the system was only possible with a student ID and from within the campus network. As it was not possible to install password managers on the lab machines and the use of personal laptops was not allowed, the use of password managers was largely avoided. If participants used a password manager on another device they would have to enter the stored password manually.

The website was used from October 2014 to April 2015 for Study 1 and from October 2015 to April 2016 for Study 2, thus for two consecutive academic years.

Password strength was calculated with the help of `zxcvbn.js` [54]. This in an open-source and JavaScript strength calculator that uses pattern matching and minimum entropy calculation. For this research, the score metric was used. It delivers a strength value between 0 and 4 that indicates whether the number of guesses required to break the password is less than 10^2 (score 0), 10^4 (score 1), 10^6 (score 2), 10^8 (score 3), or above (score 4) [1]. For example, the password "password" gets a rating of 0, where a password like "bootlegdrench42" is issued a rating of 4. Hence, the scores are not evenly spaced, the scale is exponential and the resulting data therefore ordinal. Password length was measured as the number of characters used for a password. For privacy and security reasons the participants' passwords were never transmitted unhashed: strength was calculated locally and the hashed password transmitted to the server.

Sample. All participants were students enrolled in technical courses, mainly specialising in Computer Science that used the web application for their studies. In Study 1, a total of 587 individuals registered to use the web application. Some students exercised their right to opt out, leaving 497 participants taking part in the study. In Study 2, 816 individuals registered to use the web application and created a password, of those 776 participants took part in the study.

Ethics. The study was conducted in agreement with the university's Ethics board. Participants were able to opt out of the experiment at enrolment, and about 15% did so. The school management would not allow us to contact the students to ask any questions because of the sensitivity and secrecy of passwords, and the fear that the students would interpret any communication as an indication that their passwords had been compromised. For privacy reasons we were not permitted to report any demographic information. We ensured that we used only public domain images during the course of this study.

Procedure. The participants were randomly assigned to the control condition or one of the experimental conditions by a script embedded in the enrolment web page.

They were informed that their actions were being logged and could be used for research purposes. They were presented with a consent form, allowing them to opt out of the experiment, but still benefit from use of the website. In all experimental conditions, the nudges were presented above the password entry field during enrolment and also during subsequent password creation events.

4 Results

4.1 Study 1

The data were first analyzed in terms of preconditions for statistical procedures such as sampling distribution and missing values. The descriptive statistics of Study 1 are listed in Table 2. The mean is reported as μ, the standard deviation as σ, the median as $\widetilde{x}$ and the interquartile range as IQR. Overall, the average password strength

[1] https://blogs.dropbox.com/tech/2012/04/zxcvbn-realistic-password-strength-estimation/ (accessed 28th September 2017)

NUDGE	PROMPT	COND
Control	”Choose a Password”	N0
Framing	”Choose a Secret”	N1
Expectation	Graph in Figure 2	N2
In-Group	Graph in Figure 3	N3
Expectation & Dynamic Strength	Graph in Figure 2 + Figure 4	N4
In-Group & Dynamic Strength	Graph in Figure 3 + Figure 4	N5
Social Norms		N6
Expectation & Reflection	Graph in Figure 2 + Reflection As a student, how strong do you think this password is? ○ Very Weak ○ Weak ○ OK ○ Strong ○ Very Strong ○ Unsure	N7
In-Group & Reflection	Graph in Figure 3 + Reflection As a computing science student, how strong do you think this password is? ○ Very Weak ○ Weak ○ OK ○ Strong ○ Very Strong ○ Unsure	N8

Table 1: Tested Nudges (N = Nudge Condition)

was rated with $\widetilde{x} = 1$ and IQR1 = 0, IQR3 = 3. The distribution of the password strength scores is depicted in Figure 5. The average password length was $\mu = 9.59$ ($\sigma = 3.25$) and $\widetilde{x} = 9$. The shortest password comprised 3, the longest 32 characters.

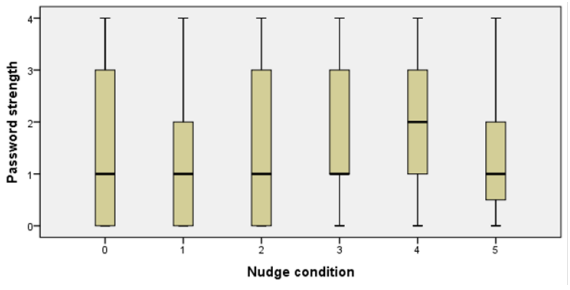

Figure 5: Password strength Study 1

Due to a non-normal sampling distribution and the password strength being measured on an ordinal scale, Mann-Whitney-U tests were conducted to compare each of the five nudge conditions with the control group. The tests were run for both password strength and length using the Benjamini-Hochberg procedure for the correction of p-values. The effect size was calculated using Cliff's Delta [13, 14] which does not make assumptions about the underlying data distribution.

Password strength in the Priming group (N1, $\widetilde{x} = 1$) did not differ significantly from the control group (N0, $\widetilde{x} = 1$), U = 3419.00, z = -.351, p = .726, Cliff's Delta = .03 [-.14, .2]. We counted two uses of the word “secret” as password in this group. However, none of the other participants, who were primed with the prompt “Provide a password” used the word ‘password’, so there is no evidence of a strong priming effect.

Likewise, there was no significant difference between the control group and the conditions In-Group Effect

		Strength Estimation					Length				
	Subjects	$\tilde{x}$	IQR1	IQR3	min	max	μ	σ	$\tilde{x}$	min	max
N0	82	1	0	3	0	4	9.46	3.83	8.00	4	32
N1	86	1	0	2	0	4	8.91	2.72	8.50	3	17
N2	83	1	0	3	0	4	9.95	3.51	9.00	6	24
N3	81	1	1	3	0	4	10.33	3.57	9.00	6	22
N4	82	2	1	3	0	4	9.76	2.53	9.00	6	17
N5	83	1	0	2	0	4	9.17	3.01	8.00	6	21

Table 2: Descriptive Statistics of user-generated passwords in Study 1 (μ = mean, σ = standard deviation, $\tilde{x}$ = median, IQR = Interquartile range).

(N3, $\tilde{x}$ = 1), U = 2955.5, z = -1.251, p = .211, Cliff's Delta = -.11 [-.28, .06], and In-Group effect with feedback (N5, $\tilde{x}$ = 1), U = 3272.5, z = -.439, p = .661, Cliff's Delta = -.04 [-.21, .13]. Finally, also the comparison between the password strength of the control group and the Expectation Effect with feedback group (N4, $\tilde{x}$ = 2) yielded an insignificant result due to the Benjamini-Hochberg adapted p-value threshold, U = 2708.00, z = -2.207, p = .027, Cliff's Delta = -.19 [-.36, -.02]. The same was true for the similar condition N2 without feedback ($\tilde{x}$ = 1), U = 3080.00, z =-1.084, p = .278, Cliff's Delta = -.09 [-.26, .08]. The effect sizes are graphically depicted in Figure 6.

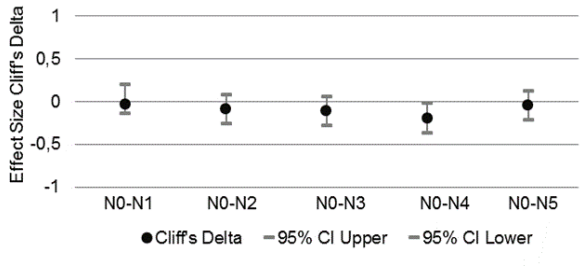

Figure 6: Effect sizes of the password strength comparisons

Password length, among others (such as use of different types of characters or of upper and lower cases), can be one factor contributing to stronger passwords. However, in line with the findings on password strength no significant effect on password length could be proven.

4.2 Study 2

The data analysis for Study 2 followed a similar approach to the one for Study 1. The descriptive statistics of Study 2 can be found in Table 3. Overall, the average password strength was rated with $\tilde{x}$ = 2, IQR1 = 0 and IQR3 = 3. The distribution of the password strength scores is shown in Figure 7. The average password length was μ = 10.02 (σ = 2.57) and $\tilde{x}$ = 9. The shortest password comprised 4, the longest 25 characters. Similar to Study 1, Kolmogorow-Smirnow tests and a visual inspection revealed deviations from a normal distribution leading to the use of nonparametric Mann-Whitney-U tests. From the original N=776 data sets, 39 had to be excluded due to technical problems with the java script strength estimator.

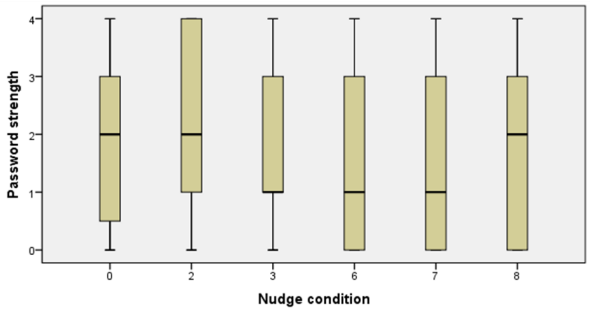

Figure 7: Password strength Study 1

Again, the control group was tested against the five experimental groups N2, N3, N6, N7 and N8 in pairwise comparisons using nonparametric Mann-Whitney-U tests and the Benjamini-Hochberg procedure for p-value correction. However, the experimental groups did not differ significantly from the control group, neither in terms of password strength nor length.

		Strength Estimation					Length				
	Subjects	$\tilde{x}$	IQR1	IQR3	min	max	μ	σ	$\tilde{x}$	min	max
N0	124	2	0.25	3	0	4	10.13	2.70	10.00	6	24
N2	124	2	1	4	0	4	10.23	2.56	10.00	6	19
N3	120	1	1	3	0	4	10.06	2.73	9.00	4	25
N6	124	2	0	3	0	4	9.80	2.42	9.00	6	16
N7	121	1	0	3	0	4	9.88	2.24	9.00	6	15
N8	124	1	0	3	0	4	10.02	2.77	9.00	5	17

Table 3: Descriptive Statistics of user-generated passwords in Study 2 (μ = mean, σ = standard deviation, $\tilde{x}$ = median, IQR = Interquartile range).

4.3 Hypothesis

Based on our findings we conclude that **H1** is not supported. The presence of the visual nudges we tested did not lead to longer and stronger passwords.

5 Discussion & Reflection

Research designs strive to maximize three criteria when collecting evidence: *generalizability, precision*, and *realism*. Since it is impossible to maximize all of these, all research designs exhibit deficiencies in one or more of these dimensions [33].

For example, survey research is generalizable whereas lab experiments are more precise, and field experiments (and case studies) are realistic while being less precise due to low controllability of confounding factors. Researchers who utilize laboratory experiments to study security behaviors can control the environment and fix a number of research variables, but realism suffers because this setting only mimics reality. Field experiments are far more realistic, but are undeniably less precise. Surveys perform poorly in terms of realism and precision.

The best research projects will probably combine the findings of surveys, lab experiments and field studies in order to offset the deficiencies of individual methods. A number of surveys have been carried out in this area [47], giving us a measure of generalizability. We contribute to the field by carrying out and reporting on our nudge-related field study, adding realism to previous findings.

After the unexpected outcome of our studies we reflected on reasons for the eight nudges seemingly making no significant impact on users' password choices. The possible explanations we considered fall into two broad categories. The first concerns potential methodological and statistical issues. The second concerns the participants: their task, aims and perceptions.

(1) Methodological considerations

The strength metric

For the purpose of our study, we decided to measure password strength with the password strength meter `zxcvbn.js`. We made this decision based on the fact that it was open-source, uses pattern matching and searches for minimum entropy. However, the score rating provided by `zxcvbn.js`, and used in our study, measures password strength on an ordinal scale with '0' indicating the number of guesses required to break the password being less than 10^2 and '4' assigned to a password requiring over 10^8 guesses. The clustering of data into 5 artificial categories, however, suppressed data variance. For example, if the number of guesses to crack a password in the control group was 1100 and that of a password in one of the experimental conditions was 9900, both passwords would be assigned a score of 2 indicating between 10^3 and 10^4 guesses required to break the password. Thus, the difference in the data would not be reflected in the score.

Although we were not aware of any alternatives when we commenced our study, there are now wrappers to run `zxcvbn.js` completely offsite. We used the open-source version of the client. To protect the participants' passwords, we did not transmit unhashed passwords — strength was calculated locally and the hashed password, together with its strength rating, transmitted to the server. The unavailability of the raw data later prevented us from calculating alternative strength estimations that might have provided a greater variance and a categorization closer to the real distribution.

The loss of information negatively affected the analysis so that it is possible that existing effects were not detected. We would therefore recommend the use of a richer classification mechanism for further studies of this

type.

Non-parametric tests

Another issue is that the ordinal password strength scale required the use of a non-parametric test. In our study, the Mann-Whitney-U test was conducted for the pairwise comparisons of the experimental and control groups. (Non-parametric tests make no assumptions about the probability distributions of the measured variables, as compared to parametric tests that require normality. Non-parametric tests are indicated where the normality requirement is violated: they are more robust against outliers and use characteristics such as the median and the central tendency to describe a distribution.)

However, if the requirements of parametric tests are met, the test power of such parametric tests is, generally speaking, higher than that of non-parametric tests. Tests with a higher test power are more likely correctly to reject a null hypothesis (no difference between groups) when the alternative hypothesis (difference between groups) is true. In our case, that means that a test with a higher test power might well have detected an existing difference between the experimental and the control groups, which the non-parametric test did not reveal. To quantify that potential impact we conducted a G*Power analysis [19] to compare the test power of an non-parametric Mann-Whitney-U test to an independent t-test. We fixed $\alpha = .05$ and sample size on 80 people per group similar to the sample sizes in study 1. We then manipulated the effect size Cohen's d required by G*Power to compare the results. We found the changes in test power to be below 2% (see Table 4).

Thus, the use of non-parametric tests might have contributed to our negative findings. Still, the analysis shows that the influence of the non-parametric vs. parametric test is rather small, whereas the influence of the effect size is much bigger. In our study 1 the effect sizes were only between .03 and .19. For future studies, it would therefore be beneficial to use a study design and password strength metric that offers greater variance and supports the deployment of parametric tests.

(2) Participant Considerations

Authentication is Complex

The focus of the experiment was solely on the password choice task. The user's specific goals and needs, in the context of the task, might not have been considered sufficiently. This is especially relevant in that security tasks are often secondary rather than primary goals [55]. Depending on the context, users aim to read mails, book a hotel or check their course details and grades. Many users might consider authentication a necessary evil that has to be overcome to reach a primary goal. It is just one among many elements in the choice-making ecosystem. Thus, it might be that the nudges tested in this study are not ineffective *per se* but that they were not powerful enough in the authentication context. For future work it would therefore be important to analyse the users' choice-making ecosystem holistically before designing a simple user interface display "intervention" to nudge users towards a change in behaviour.

Password Strength Perceptions

Studies by Ur *et al.* [48, 46] found that users' perceptions of what makes a strong password differs from the actual password security. Users succumb to several misconceptions. For example, many overestimated the security benefit of including a digit compared to other characters and underestimated the decrease in security that resulted from their use of common keyboard patterns. This might be an indication that users lack the understanding of what specifically contributes to a strong password. In the context of our results this means that the nudges might not have sufficiently enhanced the users' understanding of what makes a password stronger. Thus, feedback on password strength might be promising direction for future research.

However, the success of feedback meters in the literature, that dynamically display password strength to the user and thus constitute one form of feedback, is mixed. Studies in which users were not actively prompted to consider their password reported only marginal effects, whereas in others the meters weren't even noticed by users [7]. This confirms our earlier recommendation that future studies should engage in analyzing the targeted users, their tasks and mental models in a holistic way before designing nudges. Apart from that, one could assume that nudges which not only transport the message that passwords should be secure but also offer guidance on how to achieve this, might be more effective. This assumption, however, needs to be tested.

Password Reuse

People reuse passwords across sites [16, 25], a fact relied on by hackers globally. In a recent study by Wash and Rader [52] password re-use behaviour was investigated. The authors showed that for important accounts, such as university accounts, people re-used stronger and more complex passwords as compared to less important accounts. Thus, the difference between strong, re-used passwords (in the control group), and strong "nudged" passwords (in the experimental groups) might have been too small to detect. Apart from that, our nudges were designed to target the password creation process. If par-

			Test power	
Cohen's d	Sample size	α	independent t-test	Mann-Whitney-U test
0.1	80	0.05	0.1550283	0.1516025
0.2	80	0.05	0.3499859	0.3393193
0.3	80	0.05	0.5965318	0.5796253
0.4	80	0.05	0.8089716	0.7928030
0.5	80	0.05	0.9336887	0.9238465

Table 4: Comparative analysis of test power using the G*Power software [19].

ticipants were reusing passwords they might well have ignored the nudges altogether, rendering them impotent.

Nudges & Complex Behaviours

Nudges are targeted at at making users change a default rule or behaviour. This, however, isn't an easy task. Sunstein [42] explains that there are a number of reasons for users clinging to their default password behaviours despite the presence of nudges.

1. *First*, changing default behaviour requires active choice and effort, and the option towards which the person is being nudged might be more effortful than the default option. Nudges might be more effective where people have to choose between two options that are similar in terms of effort. In authentication, however, options are seldom similar. A stronger password increases the cost for the user in terms of time and memory load. Reusing a password is *much* less effortful than coming up with a new one.

2. *Second*, departing from the default way might be perceived to be risky and only become a realistic option if people are convinced that they should indeed change their default behaviour.

3. *Third*, people are loss averse. If the default is viewed as a reference point, a change might be considered a loss of routine or long-memorized passwords.

4. *Fourth*, password choice is cognitively expensive [20], not a simple activity. If people are already depleted for some reason they are even less likely to choose a stronger password and a visual nudge is hardly going to have the power to mitigate this.

In future studies, it would be interesting to test nudges that offers a benefit in return for the extra perceived effort. One idea suggested by Seitz, Von Zezschwitz and Hussmann [45] is to reward users with a stronger password by allowing them to keep the password longer than a weak password: applying a strength-dependent aging policy. Thus, weak passwords would be easier to type and memorise but would have to be changed more frequently whereas stronger passwords are harder to type and memorise but could be kept longer.

Limitations

As described above, this study was conducted in the field with a high degree of realism. However, field studies lack the controllability of laboratory experiments, even more so in our case where the requirements of the ethics committee constrained us in terms of collecting demographic and additional information to preserve participant privacy and anonymity.

Furthermore, the use of the password strength scores that are not evenly distributed resulted in a loss of variance and a decrease in test power. Future studies should therefore consider and compare other possibilities to quantify password strength (also see Methodological Considerations and Lessons Learned sections).

Another limitation is the limited generalisability of the sample that predominantly consisted of Computing Science students. It can therefore be expected that the sample was somewhat biased towards technically-adept, young and male participants. Another limitation concerns the design of the Figures that were presented to the participants in the experimental conditions N2 University Context, N3 School Context and the related conditions N4, N5, N7 and N8. Participants received dynamic feedback on their password strength in relation to the graph in N4 and N5. In N7 and N8 they were asked to rate the perceived strength of their passwords in relation to the graph. However, the participants in N2 and N3 did not receive feedback on their passwords. There, the nudge was intended merely to create the impression that the participants' peer group passwords ought to be

stronger than the average. In retrospect, it seems that this, on its own, did not have the power to impact password strength.

Lessons learned

A number of lessons were learned during the course of this research. We suggest the following implications that might be useful to security researchers conducting future studies:

1. **First**, the categorization of the password strength based on the `zxcvbn.js` metric used in our study resulted in a loss of information and variance of password strength. It also required the use of non-parametric tests that, generally speaking, have a lower test power than parametric tests. It is therefore possible that existing small effects were not detected. For future studies, it would be advisable to explore and compare other metrics, e.g., the exact number of guesses required to break a password.

2. **Second**, based on the literature, nudges seem to be more effective where choices are equal in terms of effort. Stronger passwords will undeniably require more effort both in terms of memory load and typing time and complexity.

3. **Third**, authentication nudges might not come into effect when users re-use passwords. Therefore, it would be interesting either to assess password reuse as a control variable, or to prevent users from reusing passwords, e.g. by applying an idiosyncratic password policy. In this case, the increased memory load would have to be acknowledged and compensated for in some way and such a policy might well introduce unanticipated and unwanted side effects.

4. **Fourth**, to better comprehend participants' understanding of secure password creation, we ought to conduct further studies exploring their mental models. It could also be useful to compare different user groups, such as laypersons and experts, who possess different levels of knowledge and perhaps engage in different decision-making strategies. Depending on the outcome of those studies, nudges that not only increase awareness, but also offer guidance on how to create stronger passwords, might be a more promising approach.

6 Conclusion

The research reported in this paper investigated the viability of a number of nudges in the authentication context. We manipulated the choice architecture to encourage the choice of stronger passwords. We discovered that password strength was not impacted by the visual nudges.

Having reflected on our findings, we were reminded of the complexity of the password creation event. It is influenced by so many more factors than the mere appearance of the surrounding user interface. We learned some valuable lessons during the course of this research and we conclude the paper by presenting a list of these to assist other researchers wishing to work in this area.

Acknowledgement

We thank our shepherd for his support in refining this paper for the LASER workshop.

We obtained ethical approval from College of Science and Engineering at the University of Glasgow to carry out nudge-related research using the website (Approval #300140006).

The research reported in this paper was supported by the German Federal Ministry of Education and Research (BMBF) and by the Hessian Ministry of Science and the Arts within CRISP (www.crisp-da.de/).

References

[1] ABDULLAH, M. D. H., ABDULLAH, A. H., ITHNIN, N., AND MAMMI, H. K. Towards identifying usability and security features of graphical password in knowledge based authentication technique. In *Modeling & Simulation, 2008. AICMS 08. Second Asia International Conference on* (2008), IEEE, pp. 396–403.

[2] ALMUHIMEDI, H., SCHAUB, F., SADEH, N., ADJERID, I., ACQUISTI, A., GLUCK, J., CRANOR, L. F., AND AGARWAL, Y. Your location has been shared 5,398 times!: A field study on mobile app privacy nudging. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (New York, NY, USA, 2015), CHI '15, ACM, pp. 787–796.

[3] BALEBAKO, R., LEON, P. G., ALMUHIMEDI, H., KELLEY, P. G., MUGAN, J., ACQUISTI, A., CRANOR, L. F., AND SADEH, N. Nudging users towards privacy on mobile devices. In *Proc. CHI 2011 Workshop on Persuasion, Nudge, Influence and Coercion* (2011), ACM.

[4] BATESON, M., CALLOW, L., HOLMES, J. R., ROCHE, M. L. R., AND NETTLE, D. Do images of 'watching eyes' induce behaviour that is more pro-social or more normative? A field experiment on littering. *Public Library of Science One 8*, 12 (2013), e82055:1–9.

[5] BHATTACHARYYA, D., RANJAN, R., ALISHEROV, F., AND CHOI, M. Biometric authentication: A review. *International Journal of u-and e-Service, Science and Technology 2*, 3 (2009), 13–28.

[6] BONNEAU, J., HERLEY, C., VAN OORSCHOT, P. C., AND STAJANO, F. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *Security and Privacy (SP), 2012 IEEE Symposium on* (2012), IEEE, pp. 553–567.

[7] BONNEAU, J., HERLEY, C., VAN OORSCHOT, P. C., AND STAJANO, F. Passwords and the evolution of imperfect authentication. *Communications of the ACM 58*, 7 (2015), 78–87.

[8] CARNAVALET, X. D. C. D., AND MANNAN, M. A large-scale evaluation of high-impact password strength meters. *ACM Transactions on Information and System Security (TISSEC) 18*, 1 (2015), 1–32.

[9] CASTANO, E., YZERBYT, V., PALADINO, M.-P., AND SACCHI, S. I belong, therefore, I exist: Ingroup identification, ingroup entitativity, and ingroup bias. *Personality and Social Psychology Bulletin 28*, 2 (2002), 135–143.

[10] CHOE, E. K., JUNG, J., LEE, B., AND FISHER, K. Nudging people away from privacy-invasive mobile apps through visual framing. In *IFIP Conference on Human-Computer Interaction* (2013), Springer, pp. 74–91.

[11] CIALDINI, R. B., AND TROST, M. R. Social influence: Social norms, conformity and compliance. In *The handbook of social psycholog*, D. T. Gilbert, S. T. Fiske, and G. Lindzey, Eds., 4 ed. McGraw-Hill, New York, 1998, pp. 151–192.

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

[13] CLIFF, N. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin 114*, 3 (1993), 494–509.

[14] CLIFF, N. Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research 31*, 3 (1996), 331–350.

[15] CRAWFORD, J. Assessing the value of formal control mechanisms on strong password selection. *International Journal of Secure Software Engineering (IJSSE) 4*, 3 (2013), 1–17.

[16] DAS, A., BONNEAU, J., CAESAR, M., BORISOV, N., AND WANG, X. The tangled web of password reuse. In *NDSS* (2014), vol. 14, pp. 23–26.

[17] EGELMAN, S., SOTIRAKOPOULOS, A., MUSLUKHOV, I., BEZNOSOV, K., AND HERLEY, C. Does my password go up to eleven?: the impact of password meters on password selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, 2013), ACM, pp. 2379–2388.

[18] FARCASIN, M., AND CHAN-TIN, E. Why we hate it: two surveys on pre-generated and expiring passwords in an academic setting. *Security and Communication Networks 8*, 13 (2015), 2361–2373.

[19] FAUL, F., ERDFELDER, E., LANG, A.-G., AND BUCHNER, A. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods 39*, 2 (2007), 175–191.

[20] GROSS, T., COOPAMOOTOO, K., AND AL-JABRI, A. Effect of cognitive depletion on password choice. In *The LASER Workshop: Learning from Authoritative Security Experiment Results (LASER 2016)* (2016), USENIX Association, pp. 55–66.

[21] HALPERN, D. *Inside the Nudge Unit: How small changes can make a big difference*. WH Allen, London, 2015.

[22] HOLDEN, J. Memorandum to the Heads of Executive Departments and Agencies. Implementation Guidance for Executive Order 13707: Using Behavioral Science Insights to Better Serve the American People, 2015. Sept 15. Executive Office of the President. Office of Science and Technology Policy `https://www.whitehouse.gov/the-press-office/2015/09/15/executive-order-using-behavioral-science-insights\-better-serve-american` Accessed 19 September 2016.

[23] HORCHER, A.-M., AND TEJAY, G. P. Building a better password: The role of cognitive load in information security training. In *International Conference on Intelligence and Security Informatics, 2009. ISI'09* (The Hague, 2009), IEEE, pp. 113–118.

[24] INGLESANT, P. G., AND SASSE, M. A. The true cost of unusable password policies: password use in the wild. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, 2010), ACM, pp. 383–392.

[25] IVES, B., WALSH, K. R., AND SCHNEIDER, H. The domino effect of password reuse. *Communications of the ACM 47*, 4 (2004), 75–78.

[26] JESKE, D., COVENTRY, L., BRIGGS, P., AND VAN MOORSEL, A. Nudging whom how: It proficiency, impulse control and secure behaviour. In *Personalizing Behavior Change Technologies CHI Workshop* (Toronto, 27 April 2014), ACM.

[27] JOSIAM, B. M., AND HOBSON, J. P. Consumer choice in context: the decoy effect in travel and tourism. *Journal of Travel Research 34*, 1 (1995), 45–50.

[28] KEITH, M., SHAO, B., AND STEINBART, P. A behavioral analysis of passphrase design and effectiveness. *Journal of the Association for Information Systems 10*, 2 (2009), 63–89.

[29] KEITH, M., SHAO, B., AND STEINBART, P. J. The usability of passphrases for authentication: An empirical field study. *International Journal of Human-Computer Studies 65*, 1 (2007), 17–28.

[30] KRITZINGER, E., AND VON SOLMS, S. H. Cyber security for home users: A new way of protection through awareness enforcement. *Computers & Security 29*, 8 (2010), 840–847.

[31] LUCK, M., AND D'INVERNO, M. Constraining autonomy through norms. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2* (Bologna, 2002), ACM, pp. 674–681.

[32] MAGNET, S. *When biometrics fail: Gender, race, and the technology of identity*. Duke University Press, Durham, USA, 2011.

[33] MCGRATH, E. Methodology matters: Doing research in the behavioral and social sciences. In *Readings in Human-Computer Interaction: Toward the Year 2000 (2nd ed*. Morgan Kaufman, 1995, pp. 152–169.

[34] OLIVER, A. Is nudge an effective public health strategy to tackle obesity? Yes. *British Medical Journal 342* (2011).

[35] ORAZI, D. C., AND PIZZETTI, M. Revisiting fear appeals: A structural re-inquiry of the protection motivation model. *International Journal of Research in Marketing 32*, 2 (2015), 223–225.

[36] RAYNER, G., AND LANG, T. Is nudge an effective public health strategy to tackle obesity? No. *British Medical Journal 342* (2011), d2168:1–2.

[37] ROSENTHAL, R., AND JACOBSON, L. *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. Holt, Rinehart & Winston, Wales, 1968.

[38] SASSE, M. A. Usability and trust in information systems. In *Cyber Trust & Prevention Project*. Edward Elgar, 2005.

[39] SHAY, R., KELLEY, P. G., KOMANDURI, S., MAZUREK, M. L., UR, B., VIDAS, T., BAUER, L., CHRISTIN, N., AND CRANOR, L. F. Correct horse battery staple: Exploring the usability of system-assigned passphrases. In *Proceedings of the Eighth Symposium on Usable Privacy and Security* (2012), ACM, pp. 7–26.

[40] SOTIRAKOPOULOS, A. *Influencing user password choice through peer pressure*. PhD thesis, The University Of British Columbia (Vancouver), 2011.

[41] SUNSTEIN, C. R. Nudges Do Not Undermine Human Agency. *Journal of Consumer Policy 38*, 3 (2015), 207–210.

[42] SUNSTEIN, C. R. Nudges that fail. *Behavioural Public Policy 1*, 1 (2017), 4–25.

[43] THALER, R. H., AND SUNSTEIN, C. R. *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press, 2008.

[44] THE BEHAVIOURAL INSIGHTS TEAM. Who we are, 2014. http://www.behaviouralinsights.co.uk/about-us/ Accessed 19 Sept, 2016.

[45] TOBIAS SEITZ, EMANUEL VON ZEZSCHWITZ, S. M., AND HUSSMANN, H. Influencing self-selected passwords through suggestions and the decoy effect. In *EuroUSEC* (Darmtsadt, 2016), Internet Society.

[46] UR, B., BEES, J., SEGRETI, S. M., BAUER, L., CHRISTIN, N., AND CRANOR, L. F. Do users' perceptions of password security match reality? In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), ACM, pp. 3748–3760.

[47] UR, B., KELLEY, P. G., KOMANDURI, S., LEE, J., MAASS, M., MAZUREK, M. L., PASSARO, T., SHAY, R., VIDAS, T., AND BAUER, L. How does your password measure up? the effect of strength meters on password creation. In *Presented as part of the 21st USENIX Security Symposium (USENIX Security 12)* (Bellevue, 2012), USENIX, pp. 65–80.

[48] UR, B., NOMA, F., BEES, J., SEGRETI, S. M., SHAY, R., BAUER, L., CHRISTIN, N., AND CRANOR, L. F. "I Added '!'at the End to Make It Secure": Observing password creation in the lab. In *Symposium on Usable Privacy and Security (SOUPS)* (2015), pp. 123–140.

[49] VANCE, A., EARGLE, D., OUIMET, K., AND STRAUB, D. Enhancing password security through interactive fear appeals: A web-based field experiment. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on* (Hawai'i, 2013), IEEE, pp. 2988–2997.

[50] VERWEIJ, M., AND HOVEN, M. V. D. Nudges in public health: paternalism is paramount. *The American Journal of Bioethics 12*, 2 (2012), 16–17.

[51] WARKENTIN, M., DAVIS, K., AND BEKKERING, E. Introducing the Check-off password system (COPS): an advancement in user authentication methods and information security. *Journal of Organizational and End User Computing (JOEUC) 16*, 3 (2004), 41–58.

[52] WASH, R., RADER, E., BERMAN, R., AND WELLMER, Z. Understanding password choices: How frequently entered passwords are re-used across websites. In *Symposium on Usable Privacy and Security (SOUPS)* (2016), pp. 175–188.

[53] WEIR, M., AGGARWAL, S., COLLINS, M., AND STERN, H. Testing metrics for password creation policies by attacking large sets of revealed passwords. In *Proceedings of the 17th ACM Conference on Computer and Communications Security* (2010), ACM, pp. 162–175.

[54] WHEELER, D. L. zxcvbn: Low-budget password strength estimation. In *USENIX Conference 2016* (Vancouver, August 2016), USENIX, pp. 157–173.

[55] WHITTEN, A., AND TYGAR, J. D. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0. In *USENIX Security Symposium* (1999), vol. 348.

[56] YEVSEYEVA, I., MORISSET, C., AND VAN MOORSEL, A. Modeling and analysis of influence power for information security decisions. *Performance Evaluation 98* (2016), 36–51.

[57] ZHANG, Y., MONROSE, F., AND REITER, M. K. The security of modern password expiration: An algorithmic framework and empirical analysis. In *Proceedings of the 17th ACM Conference on Computer and Communications Security* (2010), ACM, pp. 176–186.