COMPUTING
SCIENCE

Title: Why Johnny Cannot Remember His Password -- An Empirical Investigation

Names: Thomas Gross, Kovila Coopamootoo, Amina Al-Jabri

TECHNICAL REPORT SERIES

No. CS-TR-1509 2017

TECHNICAL REPORT SERIES

No: CS-TR-1509 Date: July 2017

Title: Why Johnny Cannot Remember His Password -- An Empirical

Authors: Thomas Gross, Kovila Coopamootoo, Amina Al-Jabri

Abstract:

Memorability vis-a-vis password strength and reuse is one of the major issues of the prevalent authentication method. The situation is aggravated by security fatigue.

We investigate how usersâ€™ password memorability differ over password reuse and strength as well as across cognitively depleted and undepleted groups.
Non-computer science students (N = 100) were randomly assigned to two groups, asked to generate a password and to return to the lab a week later to login. One group was cognitively depleted, the other was not. Password reuse and strength were observed. Password memorability was measured and compared across depletion groups, reuse and strength.

Agreeable users are more likely to create a new password (OR = 5). Men were four times as likely to create a new password compared to women. Users who have reused an existing password are more than 100 times as likely to recall their pass- word compared to users who created a new one. Users who have been cognitively depleted at the time of registration are less likely to recall their password (OR = 0.032). However, surprisingly, the likelihood to recall the password was neither significantly impacted by last time of use OR = 0.999 [0.995,1.002] nor by the password strength OR = 0.981 [0.737, 1.303].

This is the first study to establish empirically
(a) that personality traits influence whether a user reuses an existing password, (b) that cognitive depletion at time of registration negatively impacts memorability, and (c) that last time of use of a reused password and password strength does not have a significant impact on memorability.

Printed and published by Newcastle University,
Computing Science, Claremont Tower, Claremont Road,
Newcastle upon Tyne, NE1 7RU, England.

Bibliographical details

Title and Authors

NEWCASTLE UNIVERSITY
Computing Science. Technical Report Series. CS-TR- 1509

TITLE: Why Johnny Cannot Remember His Password -- An Empirical Investigation

Abstract: Memorability vis-a-vis password strength and reuse is one of the major issues of the prevalent authentication method. The situation is aggravated by security fatigue.

We investigate how usersâ€™ password memorability differ over password reuse and strength as well as across cognitively depleted and undepleted groups.
Non-computer science students (N = 100) were randomly assigned to two groups, asked to generate a password and to return to the lab a week later to login. One group was cognitively depleted, the other was not. Password reuse and strength were observed. Password memorability was measured and compared across depletion groups, reuse and strength.

Agreeable users are more likely to create a new password (OR = 5). Men were four times as likely to create a new password compared to women. Users who have reused an existing password are more than 100 times as likely to recall their pass- word compared to users who created a new one. Users who have been cognitively depleted at the time of registration are less likely to recall their password (OR = 0.032). However, surprisingly, the likelihood to recall the password was neither significantly impacted by last time of use OR = 0.999 [0.995,1.002] nor by the password strength OR = 0.981 [0.737, 1.303].

This is the first study to establish empirically
(a) that personality traits influence whether a user reuses an existing password, (b) that cognitive depletion at time of registration negatively impacts memorability, and (c) that last time of use of a reused password and password strength does not have a significant impact on memorability.

About the authors: Current Research: Cyber Security, Privacy and Evidence-based Methods for Security

Thomas Gross

I'm a Tenured Reader in System Security (Associate Professor) at the Newcastle University. I’m the Director of the Centre for Cybercrime and Computer Security (CCCS), a UK Academic Centre of Excellence in Cyber

Security Research (ACE-CSR). I'm a member of the Secure and Resilient Systems group and the Centre for Software Reliability (CSR).

Dr Kovila Coopamootoo: is currently a Research Associate in the Secure & Resilient Systems group, School of Computing Science at Newcastle University. Her research involves cognitive effort in decision-making (using eye-tracking and physiological measurements) and mental models (including mixed method research). Her interests expand to usable privacy and security and user decision-making under uncertainty. She is currently supported by the FutureID project. Previously she was involved in the "Hyper-privacy: Case of Domestic Violence (Hyper-DoVe)" project in applying technologies to enable survivors of domestic violence to look for help while protecting their privacy. In particular, she was involved in refining and evaluating a toolkit of privacy technologies that enable survivors to achieve privacy while accessing information online, with minimum effort and without leaving digital record of their visit. The work was carried out in collaboration with the Angelou Centre.

Amina Salim Mohd Al-Jabri was a MSc student of Newcastle University in the group Secure and Resilient Systems (SRS) working on the reported experiment on the effect of cognitive effort and depletion and password choice as part of her MSc thesis.

Suggested keywords:

password strength, memorability, personality traits, evidence-based, empirical, experiment

Why Johnny Cannot Remember His Password
An Empirical Investigation

Thomas Groß
Newcastle University

Kovila Coopamootoo
Newcastle University

Amina Al-Jabri
Newcastle University

Abstract

Memorability vis-à-vis password strength and reuse [46, 47, 31, 44, 13] is one of the major issues of the prevalent authentication method. The situation is aggravated by security fatigue [14, 35].

We investigate how users' password memorability differ over password reuse and strength as well as across cognitively depleted and undepleted groups.

Non-computer science students ($N = 100$) were randomly assigned to two groups, asked to generate a password and to return to the lab a week later to login. One group was cognitively depleted, the other was not. Password reuse and strength were observed. Password memorability was measured and compared across depletion groups, reuse and strength.

Agreeable users are more likely to create a new password ($OR = 5$). Men were four times as likely to create a new password compared to women. Users who have reused an existing password are more than 100 times as likely to recall their password compared to users who created a new one. Users who have been cognitively depleted at the time of registration are less likely to recall their password ($OR = 0.032$). However, surprisingly, the likelihood to recall the password was neither significantly impacted by last time of use $OR = 0.999$ $[0.995, 1.002]$ nor by the password strength $OR = 0.981$ $[0.737, 1.303]$.

This is the first study to establish empirically (a) that personality traits influence whether a user reuses an existing password, (b) that cognitive depletion at time of registration negatively impacts memorability, and (c) that last time of use of a reused password and password strength does not have a significant impact on memorability.

1 Introduction

How well users remember their passwords is a major issue of the still prevalent authentication method. Sasse et al. [31] discussed the difficulty of users to remember passwords as part of the 'weakest link,' an area where usability and effective security go astray. Smith [33], for instance, has decried strong passwords as inherently impossible to remember. Consequently, there have been research efforts that sought to afford users strong passwords that are at the same time memorable, for instance, through meaningful, pronounceable [46] or mnemonic [44] passwords.

Even if these approaches benefit from strength as well as memorability, large-scale empirical evaluation showed that the user's password habits follow different routes [13]. For instance, users were found to have a limited set of passwords that they reuse across multiple sites. Clearly, their habit to reuse passwords would impact memorability. However, even if we consider that—on average—a user is said to have 6.5 passwords, each shared across 3.9 different sites, the question remains: How do

the users' habits differ depending on their personality traits and their current cognitive state? How do the users' choices impact memorability?

We set out to answer these questions in an empirical experiment:

1. RQ-R *How do personality traits and cognitive depletion impact the likelihood of a user to create a completely new password or to reuse an existing one?*
2. RQ-M *How do cognitive depletion as well as the users' choices on reuse and password strength influence the likelihood to remember the password?*

We induce cognitive depletion [4] in an experiment with $N = 100$ participants and observe personality traits as well as password strategies of the users. Cognitive depletion is especially interesting because we believe it to contribute to security fatigue [14, 35] as a short-term factor. We measure the user's choice to create a completely new password as well as the capacity to remember the password after one week.

Contribution This study is the first one to propose an experiment investigating the impact personality traits and cognitive depletion on password reuse and memorability systematically. It is, thereby, shedding light on what drives the user's password habits that been observed macroscopically and what factors are further impacting the likelihood to remember down the line.

It shows that password reuse behavior differs across gender, the Big Five personality traits, and especially agreeableness. The experiment shows further that the creation of a completely new password severely impacts the likelihood to remember and that cognitive depletion during the registration makes matters worse. Finally, we see that the impact of password strength measured with zxcvbn is actually quite limited and less than the impact of either new-password creation or cognitive depletion.

Outline After having introduced the background of this research, Section 3 introduces the method of this work, including the operationalization of the aim, procedure, and manipulation and measurement methods. Section 4 contains the results of the research, especially two logistic regressions: one for the likelihood of a user to choose a completely new password and one for the likelihood of a user to remember the password one week after registration. Section 5 discusses the results and interprets emergent themes.

2 Background

This section looks at literature on password strengths in relation to users' ability to remember them. We then introduce cognitive effort and explain the state of ego depletion. Lastly we review how personality traits influence decisions.

2.1 Password Memorability

Password memorability has been a key issue in authentication and security for decades [46, 47, 31, 44, 13]. Characteristics of human memory such as the limited working memory capacity and that memory decays over time pertain to passwords, as summarised by [31]. The paradox between password strength and memorability has been in the community for some time with the idea that best practices imply that "the password must be impossible to remember" [33]. In addition it has been suggested that recalling strong passwords is an impossible task, with the reasoning pointing towards non-meaningful passwords [29, 32]. This issue has been used as motivation in the research community such as by [10].

There are however examples of research investigating solutions to this paradox in specific ways. For example, a case has been made for strong passwords that are also meaningful, hence aiding recall [31]. Other research found that particular password strategies such as those composed from a mnemonic are as strong as random passwords and are less difficult to be remembered [44], that the character format of passwords matters for recall and that pronounceable passwords have better recall results [46]. In addition, that memorability of text passwords can be aided with question-and-answer based on user's perceptions, personal interests and

personal history (cognitive passwords) [45], a predefined list of cues and responses unique to the user (associative passwords) [46] and codes with semantic meaning ("pass-sentences") [34].

However while the memorability of particular password strategies have been investigated, the field still lacks research on the combination of strength and memorability of passwords that are chosen freely by the user.

2.2 Password Habits

It is believed that when forced to comply to security policies such as monthly password reset, a large number of users are frustrated [19]. Users hence develop a variety of password strategies. The strategies include writing passwords down, incrementing the number in the password at each reset [2], storing passwords in electronic files and reusing or recycling old passwords [19].

Users also have a variety of password habits. Research employing self-report had users claiming to have at most three passwords and reusing passwords twice [15]. However, an observational study found that the average user has 6.5 passwords, each shared across 3.9 different sites and that each user has 25 accounts requiring passwords and type 8 passwords per day [13]. The number of sites per password is thought to increase with age of client, with weak passwords being shared at more sites. The number of logins and password strength is dependent on the nature of the site. Longer passwords are composed of digits or alphanumeric characters [13]. Das et al.[9] offer a further comprehensive study of password reuse, including an empirical estimate of the rate of guessing at 43%.

2.3 Cognitive Effort and Depletion

Human beings have a limited store of cognitive energy or capacity [4] with self-control tasks, choice and decision-making drawing from this inner resource. As a muscle that gets tired with exertion, self-control tasks cause short-term impairments in subsequent self-control tasks. This is termed a state of *ego depletion* or *cognitive depletion*. There are levels of depletion beyond which individuals may be unable to control themselves effectively, regardless of what is at stake [5] and in unrelated sphere of activity [4].

The impact of cognitive effort and depletion has been studied by Groß, Coopamootoo and Al-Jabri [17], finding that cognitive depletion has a statistically significant effect on password strength.

2.4 Personality Traits

Personality refers to "individual differences in characteristic patterns of thinking, feeling and behaving" [1] of the human. The *Big Five* [16, 21, 23] is a general taxonomy spanning across five dimensions that provides a way to model personality. The five dimensions of personality are openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism.

Personality traits are thought to positively or negatively impact decisions, as exemplified in the health domain [26] or to have an impact on the propensity for risky decisions [25]. In security research, specific personality traits have been linked to different contexts. For instance impulsive individuals are more likely to fall for phishing mails, agreeable individuals likely to choose stronger passwords and individuals are susceptible to social engineering attacks depending on their personality traits [39].

2.5 Demographics

Age, gender and education have previously been observed to impact password choice. Bonneau [6] found that there is a split-effect in online/offline guessability between men and women. Further, he reported a trend towards stronger passwords with age. Mazurek et al. [27] have found that men produced slightly stronger passwords than women. They also found differences by colleges, that is, type of education.

[1] `http://www.apa.org/topics/personality`

3 Method

3.1 Aim

RQ-R *How do personality traits and cognitive depletion at registration time impact the user's choice to create a new password or reuse an existing password?*

- We induce the independent variable (IV) cognitive depletion.
- We observe the covariate variables Big Five personality traits and demographics.
- We measure the user's choice on how the password is chosen in a self-report questionnaire as dependent variable (DV), especially whether a new password is created or an existing one reused.

Figure 1a gives an overview of the relations of the predictors to the DV.

We remark here that it is essential for the design to leave it to the user to make a choice to create or reuse a password freely. If we were to induce this behavior as an independent variable, we expose the experiment to the experimenter expectancy effect [30] which can be amplified by personality traits [18]. This way the experimenter is not invested either way in the users creating a new password or reusing an existing one as they see fit.

We asked participants about the strategy they have employed for the particular password just created. Due to its specificity, we expect this to be a more accurate statement than self-report statements on password habits, observed to deviate from observational studies [15, 13].

RQ-M *How do cognitive depletion at registration time, the decision to create or reuse a password, time of prior use and password strength impact the likelihood to recall a password?*

- We induce the independent variable (IV) cognitive depletion.
- We observe the user's choice to create a new password or to reuse an existing one.
- We further observe the last time a reused password was used.
- We measure as dependent variable (DV), whether the user is able to login with the chosen password after one week.

Figure 1b gives relates the predictors to the DV.

We take into account the *time of prior use* instead of the frequency of the password use because self-report statements on reuse frequency have been shown to be inaccurate. Furthermore, the frequency estimates are bound to be impacted by the availability bias [38, 41]: Users tend to use the ease of recall as a heuristic for frequency estimates.

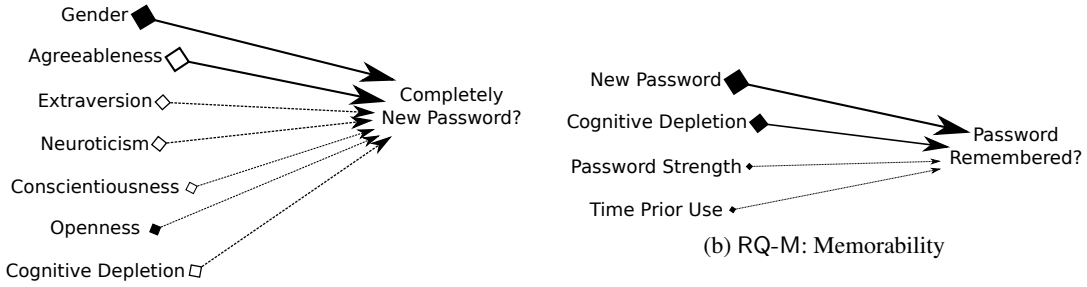
3.2 Participants

The sample consisted of university students, $N = 100$, of which 50 were women. The mean age was 28.18 years ($SD = 5.241$) for the 83 participants who revealed their age. The participants were balanced by gender and assigned randomly to either the depletion ($n = 50$) or control ($n = 50$) condition. They were mostly non-computer science students, of mainly international background.

Our demographics questionnaire included a 10-item security awareness questionnaire, which was based on the security awareness metrics of SANS Securing the Human[2]. Men scored statistically significantly higher ($M = 5.64, SD = 1.99$) than women ($4.15, SD = 1.86$), $t(93) = 3.78, p < .001$ with $\Delta M = 1.49$ $[0.71, 2.28]$ ($d = 0.75$). After the random assignment to experiment and control condition the difference in security awareness between groups was not statistically significant, $t(93) = -1, p = .3$.

Tiredness and cognitive depletion over the course of a day are affected by the participants' circadian rhythm. Hence to control the confounds of the circadian rhythm, the experiment runs were balanced in time-of-day for depletion ($M = 4.167$, $SD = 1.403$) and control ($M = 4.167$, $SD = 1.642$) conditions. We ran a Wilcoxon signed-rank test on the two conditions matched by time of day. We find that the distribution of participants across the two groups was not statistically different, with $Z = 0.00$ and $p = 1.00$.

[2] `https://securingthehuman.sans.org/resources/metrics`



(a) RQ-R: Password Reuse

(b) RQ-M: Memorability

Figure 1: Predictors and dependent variables of the two research questions.
An empty diamond ◇ denotes a positive effect. A filled diamond ♦ denotes a negative effect. Arrows are roughly scaled to reflect the effect size seen in the experiment and dashed for non-significant predictors.

3.3 Procedure

The experiment was designed to observe the influence of long-term factors, such as personality traits, and to compare password reuse and memorability across groups induced with cognitive depletion or not.

The experiment group was artificially cognitively depleted with tasks that required impulse control while the control group was not depleted, completing non-depleting tasks with similar length and flavor.

The procedure consisted of (a) pre-task questionnaires for demographics and personality traits, (b) a manipulation to induce cognitive depletion, (c) a manipulation check on the level of depletion, (d) a password entry for a mock-up GMail registration, and (e) a debriefing and memorability check one week after the task with a GMail login mockup. (f) a debrief questionnaires on password strategies. Figure 2 depicts the experiment design.

3.3.1 GMail Registration Task

Participants were asked to generate a new password for a Google Mail (GMail) account, on a mock-up page which was visually identical to a GMail registration. The participants were told (a) to create the account carefully and fill in all the fields; (b) to give correct and valid information; (c) that the account is highly important; and (d) that they should ensure they can remember the password. Participants were also asked to return to the lab one week after the registration task. Registered e-mail address and password were recorded. The strength of the password was measured.

3.3.2 Inducing Cognitive Depletion

We induce cognitive depletion for the experiment condition, reproducing manipulation components of Baumeister et al. [37]. In the experiment condition, the participants are asked to suppress thoughts, control impulses to follow a learned routine and to execute a cognitively effortful Stroop task. In the control condition, the participants fulfil tasks with a similar structure, flavor and length, however without the depleting conditions.

We control the strength of the manipulation with a manipulation check based on a brief mood inventory (Section 3.4.2) evaluated in the Results Section 4.1.

1. Thought suppression task In the experiment condition, the participants are shown photo of a white bear and asked *not* to think of the white bear, a procedure following Wegner et al. [42]. They are to raise their hand should they have thought of the white bear and failed to suppress the thought. In the control condition, the participants are asked to record whenever they think about a white bear.

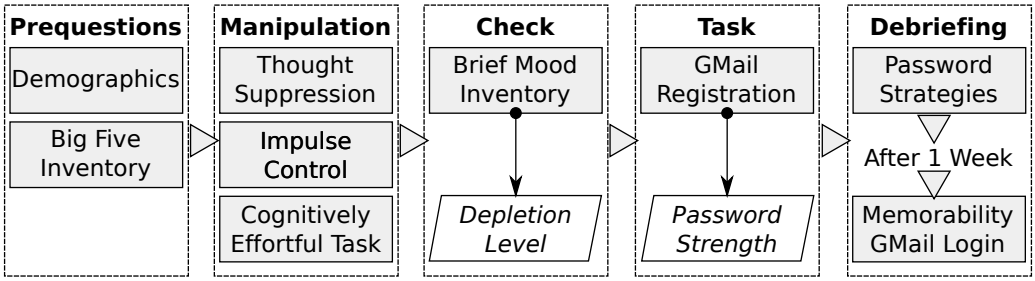


Figure 2: Overview of the experiment procedure.

However, they are not depleted because they not instructed suppress their thoughts.

2. Impulse control task This task is adapted from Muraven et al. [28]. Participants are asked to cross out all letters 'e' in a complex statistical text for five minutes. This establishes a learned routine. Then the participants are given another statistical text. In the experiment condition, the participants are asked to follow a new rule, to cross out all letters 'e' unless they are adjacent to a vowel. This rule interferes with the learned routine and asks the participants to exercise impulse control on it, which is depleting. In the control condition, the participants are asked to follow the same routine, which is non-depleting.

3. Cognitively effortful task We used the Stroop task [36] as cognitively effortful task. Participants are asked to voice the printed color of a color word. The Stroop condition is that the name of a color (e.g., 'red') is printed in a color not denoted by the name (incongruent color and name). This task is a cognitively effortful when the Stroop condition is fulfilled. The experiment condition involved answering 10 Stroop items with the Stroop condition. The control condition involved answering 10 items without Stroop condition.

3.4 Measures

Following the structure of the experiment displayed in Figure 2, we introduce the measures (a) Big Five Inventory, (b) depletion level, (c) password strength, (d) password strategies, and (e) memorability.

3.4.1 Big Five Inventory

The personality traits of the users were queried with a 60-item Berkeley Big Five Inventory (BFI) [16, 21, 23]. The inventory measures the traits (a) Openness to experience, (b) Conscientiousness, (c) Extraversion, (d) Agreeableness, and (e) Neuroticism, with a 5-point Likert-type items between Disagree strongly and Agree strongly computing the scores as means of items for each domain.

3.4.2 Manipulation Check: Depletion

We used a short form of a brief mood inventory proposed as manipulation check by Baumeister [37] rated on 5-point Likert-type items between Disagree strongly and Agree strongly, with Neither agree nor disagree as central point. Baumeister el al. [37] found that tiredness and feeling worn out are significantly affected by cognitive depletion and can therefore be used as self-report manipulation check.

3.4.3 Password Strength

We used multiple strength measures and compared their results. The list consists of

(a) the Dropbox zxcvbn ($\log_{10}$ guesses) [43],
(b) the CMU Password Guessability Service (PGS) ($\log_{10}$ guesses) [24],
(c) a Web password meter[3], with penalties for dictionary words and parts of the username (linear scale between -100 and 150),
(d) the NIST password entropy estimate [7].

[3] `http://www.passwordmeter.com`

Table 1: Correlation Matrix Password Strength Pearson correlation coefficient, all $p < .001$.

	zxcvbn	PGS	Pwd'meter	Entropy
zxcvbn	—			
PGS	0.58	—		
Pwd'meter	0.43	0.46	—	
Entropy	0.45	0.57	0.50	—

Table 1 outlines the correlations between the different tools, where we observe the highest correlations between zxcvbn and PGS. We settled on zxcvbn [43] as reference password strength meter of choice. First, it has seen adoption in the security community. Second, zxcvbn results are reproducible in that particular version of the algorithm and supporting datasets can be committed to. Third, zxcvbn protects the participants' privacy, because the metrics can be computed offline.

We stress that all analysis results stay valid irrespective of the password strength tool used.

3.4.4 Password Strategies

We asked participants about the strategies they have employed to choose their password.

1. Is this a completely new password?
2. Have you used this password before?
3. When did you use the password before?
4. Have you used a similar password before?
5. What strategy did you implement in creating a password? [freeform text]

Definition of Reuse If a participant answers the question *1. Is this a completely new password?* with yes, we call the value given a *new password*. We defined *reused password* as the complement of *new password*. Questions 2 and 4 are meant to check consistency of the user's statements and to qualify adaptation of previous password.

Last Time of Use Last time of use was measured in number of days.

In the experiment design, we opted for last time of use instead of frequency. Successful recall of information is influenced, among other things, by the number of recall repetitions as well as the last time the information was recalled. A piece of information recalled more frequently will benefit from longer memory retention. However, once the piece of information is not recalled any longer, the memory retention will diminish over time. Consequently, a password that has been used frequently in the past, but then fallen to disuse, is likely to be forgotten.

Frequency estimates, however, are known to be inaccurate. First, self-reports of users on password use frequency in self-reports did not agree with observation studies [15, 13]. Second, frequency estimates have been shown to be affected by the availability heuristic: humans tend to rely on their ease to recall the information to estimate the frequency [38, 41]. While this general tendency was reconfirmed experimentally, exceptions to that rule were found when the participants made a special cognitive effort to make their frequency estimate as accurate as possible. [1]. Thereby we expect frequency estimates to differ between undepleted and depleted states.

Furthermore, for passwords that have been reused with high frequency and are still in use, the prior time of use is an accurate proxy. For passwords that have not been reused for a longer time, the prior time of use is more accurate than the frequency, because of the diminishing memory retention.

3.4.5 Memorability

The participants were asked to return to the lab a week after the initial registration on a GMail mockup. They were then presented with a mockup GMail login screen. We measured whether the participants were able to recall their registered password (pass) or not (fail). The participants were allowed a maximum of five login attempts.

4 Results

All inferential statistics are computed with two-tailed tests and at a significance level of $\alpha = .05$.

4.1 Manipulation Check

A comparison across groups on tired and worn out suggested that the manipulation was successful (Mann-Whitney U, two-tailed, tired: $U = 368, Z = -6.299$, significance $p < .001$; worn out: $U = 669$, $Z = -4.145$, significance $p < .001$). As expected following Baumeister et al. [37] in the use of the brief mood inventory: the moods of feeling tired and feeling worn out were found to be significantly higher in the depleted group than in the control group. The effect size of the manipulation for reporting feeling tired is large ($r = 0.63$) and for feeling being worn out is medium to large ($r = 0.42$).

The tiredness score is grouped in three homogeneous groups, which we call the *depletion level*:

(a) non-depleted (disagree strongly, disagree slightly and neither agree nor disagree).
(b) effortful (agree slightly), and
(c) depleted (agree strongly).

Of the control group, 49 participants were rated non-depleted; 0 participants were rated as effortful; 1 participant was rated as depleted. Of the experiment group, 23 participants were rated non-depleted; 17 participants were rated as effortful; 10 participants were rated as depleted.

4.2 Reuse of an Existing Password

We establish to what the choice of a new vs. an existing password is impacted by personality traits and cognitive depletion.

4.2.1 Descriptive Statistics

We offer the descriptive statistics in a contingency table in Table 2. A two-tailed Fischer's exact test did not yield a significant result ($p = .215$), by which we accept the null hypothesis that the relative proportions of the depletion level variable are independent of the new password variable.

4.2.2 Logistic Regression

A logistic regression was conducted to predict the likelihood whether the participant would select a new password or reuse an old password. The predictor variables were the participant's gender, depletion level (3 categories with undepleted as baseline) and five BFI personality traits.

Table 2: Contingency table depletion level vs. whether a new password was chosen.

	New password		
Depletion level	NO	YES	Total
Undepleted	44 (61%)	28 (39%)	72
Effortful	14 (82%)	3 (18%)	17
Depleted	6 (54%)	5 (46%)	11
Total	64 (64%)	36 (36%)	100

We selected these predictors based on a number of prior experiments, which tested the impact of Big Five personality traits and cognitive effort on password choice (lab, online, and MTurk). Cognitive depletion has been shown previously to significantly impact password strength; it was our first choice of as predictor here. Gender was added to control for differences across gender with respect to personality traits. In previous results, we have seen significant impact of agreeableness, extraversion and neuroticism on different aspects of the password choice. We decided to input all BFI factors for completeness.

The null hypothesis $\mathsf{H}_{\mathsf{R},0}$ is: *None of the predictors named has a significant impact on the likelihood to create a new password.* Conversely, the alternative hypothesis $\mathsf{H}_{\mathsf{R},1}$ states: *At least one of the predictors impacts the likelihood to create a new password.*

The test of the full model in comparison to the model with the intercept only was statistically significant, Maximum Likelihood Test $\chi^2(8,100) = 25.852, p = .001$, Wald Test $\chi^2(8,100) = 17.816, p = .023$. Hence, the model is deemed able to distinguish participants reusing old passwords versus participants creating new passwords. The model explained between 20% (Hosmer & Lemeshow) and 31% (Nagelkerke) of the variance.

The model correctly classified 72% of the cases. It correctly classified participants who reused an old

password 84.4.% of all cases and participants who created a new password 50% of all cases.

Table 3 contains an overview of the coefficients and odds ratios of the logistic regression. At a significance level of .05, gender and agreeableness had statistically significant partial effects. Holding all other variables constant, women are roughly one fourth as likely as men to choose a new password.

Holding all other variables constant, agreeable participants are more likely to choose a new password. A one-point increase on a 5-point BFI Agreeableness scale is associated with the odds of choosing a new password increasing by a multiplicative factor of 5.

We plot the log odds and likelihoods for the significant predictors gender and agreeableness in Figure 3. Depletion level and other Big Five personality traits were not statistically significant.

Based on the results of the logistic regression, we reject the null hypothesis $\mathsf{H}_{\mathsf{R},0}$.

4.3 Memorability

We establish to what extent depletion level and the choice of a new vs. an existing password impacts memorability.

4.3.1 Descriptive Statistics

One of our primary interests for this study is the memorability compared across induced depletion levels. Table 4 contains the contingency table of the memorability. The proportions of this contingency table are not statistically independent, $p = .044$ (FET).

4.3.2 Logistic Regression

We computed a logistic regression modelling whether a participant could remember a password for a GMail login page one week after the registration. As predictors, the regression used the the depletion level, whether a new password was created or an old one reused, the last time of prior use of that password and the password strength in zxcvbn $\log_{10}$ number of guesses.

The predictors were chosen based on the rationale that depletion can impair working memory, which would in turn impair making memory persistent and that new passwords are less likely to be recalled due to the missing repetitions. Password strength and prior time of use were considered in the community discourse as factors for memorability. We considered personality traits and gender a long shot for memorability after a week's time and decided not to include them. Preciously, we tested a second model with the same predictors plus the condition. We removed the condition however to avoid variance inflation.

The null hypothesis $\mathsf{H}_{\mathsf{M},0}$ is stated as: *None of the predictors impacts the likelihood to recall the registered password after one week.* Conversely, the alternative hypothesis $\mathsf{H}_{\mathsf{M},1}$ is: *At least one of the predictors influences the likelihood to recall the password after one week.*

A statistically significant logistic regression equation was found, Likelihood Ratio Test $\chi^2(1, N = 100) = 51.318, p < .001$, Wald Test $\chi^2(1, N = 100) = 16.334, p = .006$. The model explained between 48% (Hosmer & Lemeshow) and 61% (Nagelkerke) of the variance.

The logistic regression classified failed recall attempts 56.5% correctly and successful recall attempts 94.8% correctly, yielding an overall accuracy of 86%.

Table 5 contains the regression coefficients for this analysis. The depletion level depleted and having chosen a new password had statistically significant effects. We reject the null hypothesis $\mathsf{H}_{\mathsf{M},0}$.

Holding all other variables constant, participants who reuse an existing password are more than 100 times as likely to remember their password compared to participants who choose a new password.

Holding all other variables constant, participants being depleted at registration time are one twentieth as likely to remember their password.

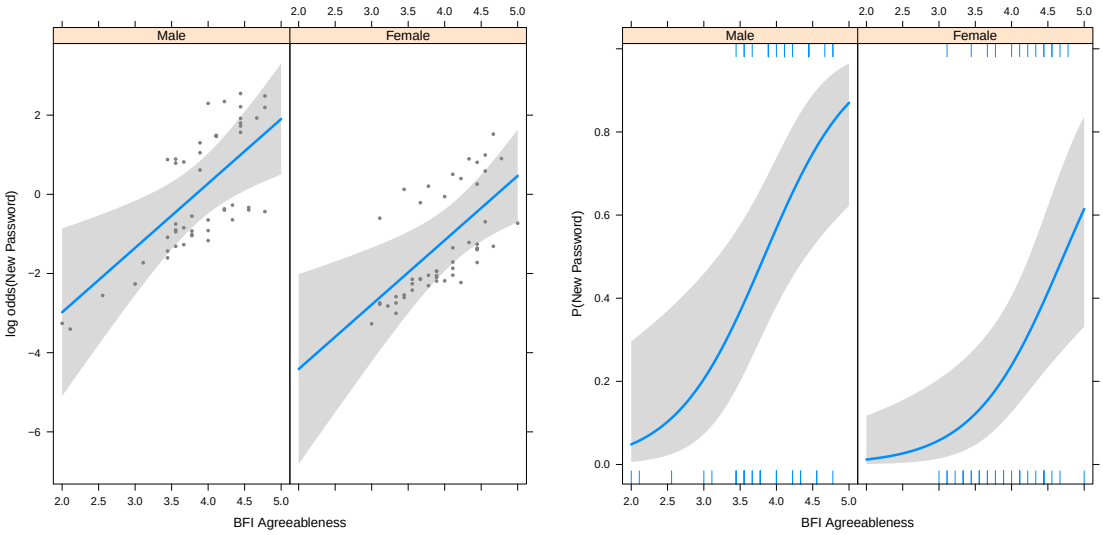
We depict the influence of the two significant predictors in Figure 4, showing the coefficients in comparison in Figure 4a and giving the likelihoods in Figure 4b.

We observe that the neither the time of prior use nor the password strength measured in zx-

Table 3: Coefficients of the logistic regression on new vs. old passwords.

Predictor		*B*	SE	Wald χ^2	df	Sig.	Odds Ratio	95% CI		*R*
								LL	UL	
Gender		−1.435	.555	6.686	1	**.01****	**0.238**	0.080	0.707	.19
Depletion	undepleted			5.305	2	.07				
	effortful	−1.497	.816	3.362	1	.067	0.224	0.045	1.109	
	depleted	.964	.778	1.536	1	.215	2.622	0.571	12.033	
BFI Agreeableness		1.627	.538	9.157	1	**.002****	**5.089**	1.774	14.597	.23
BFI Neuroticism		.617	.387	2.542	1	.111	1.853	0.868	3.954	
BFI Conscientiousness		.470	.512	.843	1	.359	1.6	0.587	4.364	
BFI Extraversion		.991	.556	3.181	1	.074	2.694	0.907	8.007	
BFI Openness		−.517	.547	.894	1	.344	.596	0.204	1.742	
Constant		−11.160	3.515	10.078	1	.002**	< .001			

Note: $R^2 = .2$ (Hosmer & Lemeshow) .23 (Cox & Snell) .31 (Nagelkerke). Model $\chi^2(9) = 25.976, p = .002$.



(a) Log odds

(b) Likelihood

Figure 3: Log odds and likelihood to create a new password from predictors gender and agreeableness. 3a shows the coefficients for agreeableness in two panels for gender. 3b displays the likelihood to create a new password when other predictors are held constant.

Table 4: Contingency table depletion level vs. whether the password was remembered.

	Remembered		
Depletion level	NO	YES	Total
Undepleted	14 (19%)	58 (81%)	72
Effortful	3 (18%)	14 (82%)	17
Depleted	6 (54%)	5 (46%)	11
Total	23 (23%)	77 (77%)	100

cvbn $\log_{10}$ guesses have a significant impact on memorability. The 95% confidence interval on the odds ratio of password strength is bracketed around 0.89 $[0.737, 1.303]$. For each increase in order of magnitude of the #guesses, the likelihood of remembering shrink by a factor of 0.89.

Figure 4 displays the log odds and likelihoods for the two significant predictors.

4.4 Effect Size of Password Strength

We have conducted logistic regressions that use different metrics for password strength ($\log_{10}$ guesses in zxcvbn and PGS, `passwordmeter.com`, NIST entropy). The regressions agreed on the limited impact of password strength on memorability irrespective of the metric used. How can we gain confidence in the effect size reported?

The reported logistic regression offers a 95% confidence interval of [0.737, 1.303] on the odds ratio: if the experiment were repeated many times, the confidence interval would contain the true population odds ratio 95% of the cases. Does the result depend on the metric used?

We compare the outcomes of zxcvbn and PGS in a meta-analysis and estimate the effect size across metrics. Both tools measure guessability in $\log_{10}$ guesses and are thereby well comparable. Figure 5 on page 13 shows the forest plot for the analysis. According to the fixed-effect model, the effect size 0.93 $[0.77, 1.12]$ offers a more robust estimate of the impact of password guessability on memorability. By this estimate, we expect the memorability after one week to decrease by a factor of 0.93 for each increase in one point on a $\log_{10}$ guessability scale.

4.5 Model Properties

4.5.1 Accuracy

Password Reuse There were four cases with large residuals (4%), with values greater than 2 or less than -2. All of these case have a Cook's distance well less than 1 and, thereby, no undue influence in the model. There were six cases with more than twice the average leverage, however, none of the cases had more than three times the average leverage. For the assumption of independence, the Durbin-Watson statistic was $2.4, p = .032$, which is still in the bounds Field recommends [12, p. 921]. In terms of multicollinearity, average Variance Inflation Factor (VIF) was small: 1.2. All tolerances were well above .5. Hence, we conclude that there was no collinearity in the data. Finally, we assess the residuals. Considering the QQ Normal plot in Figure 6a on page 13, we observe deviations from normality below 0. All things considered, we conclude the model to be sufficiently accurate.

Memorability There were three cases with large residuals (3%), with residuals greater than 2 or less than -2. The Cook's distance was well below 0.25 alleviating any concerns. Of the 17 cases with more than double the average leverage, two had more than three times the average leverage. Durbin-Watson confirmed the assumption of independence with a test statistic of $2.14, p = .51$. We reject the the hypothesis that there is multicollinearity in the data because the average Variance Inflation Factor (VIF) is sufficiently small: 1.39 and because all tolerances are well above 0.2. From the QQ Normal plot in Figure 6b on page 13, we observe deviations from normality of the residuals especially around 0. While the distribution of residues is not fully normal, we still perceive the model as sufficiently accurate.

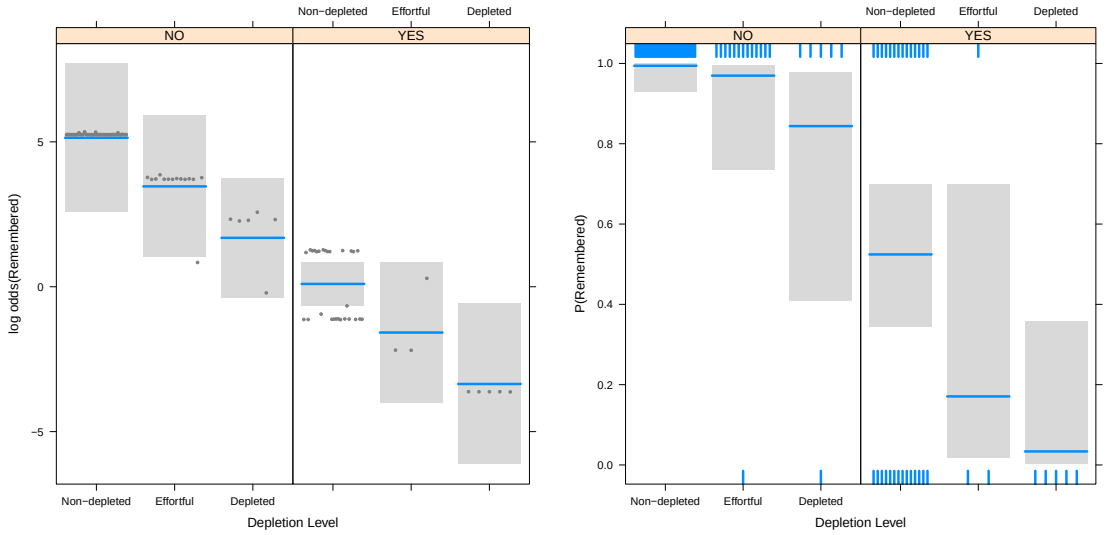
4.5.2 Prediction Performance

Whereas our models are focused on causal analysis, we evaluate the prediction performance of the models as well on the given sample. Figure 7 on page 13 contains the Receiver Operating Charac-

Table 5: Coefficients of the logistic regression memorability with $\log_{10}$ password guesses from zxcvbn.

Predictor	B	SE	Wald χ^2	df	Sig.	Odds Ratio	95% CI LL	95% CI UL	R
Depletion									
effortful	−1.678	1.265	1.8	1	.18	0.187	0.008	1.808	
depleted	−3.454	1.423	5.9	1	**.015***	**0.032**	0.001	0.349	.19
New Password	−5.044	1.282	15.5	1	**< .001*****	**0.006**	0.001	0.048	.35
Time Prior Use	−.001	.002	0.55	1	.46	0.999	0.995	1.002	
Zxcvbn $\log_{10}$ Guesses	−.019	.141	.018	1	.89	0.981	0.737	1.303	
Constant	5.268	1.738	9.2	1	.002**	194			

Note: $R^2 = .476$ (Hosmer & Lemeshow) .4 (Cox & Snell) .61 (Nagelkerke). Model $\chi^2(6) = 51.318, p < .001$.



(a) Log odds

(b) Likelihood

Figure 4: Log odds and likelihood to remember a password from whether a new password was chosen and depletion level at registration time. 4a shows the coefficients for depletion levels in two panels for reused and new passwords. 4b shows the likelihood to remember when other predictors are held constant.

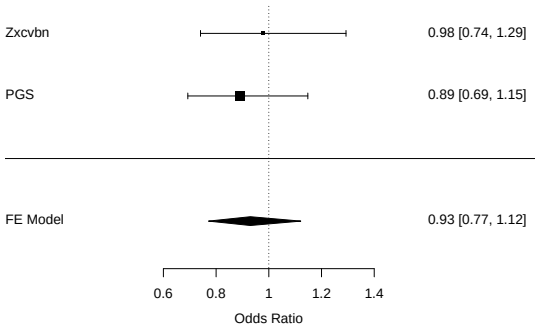
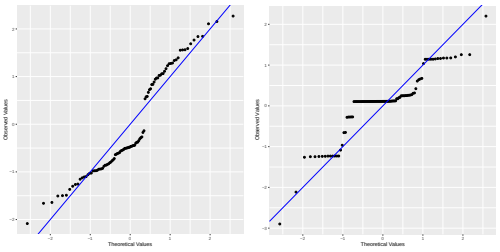


Figure 5: Forest plot of the effect of password strength in $\log_{10}$ guesses on memorability.



(a) Reuse (b) Memorability

Figure 6: QQ-Plots for both models.

teristic (ROC) curve of the both models, reuse in a dashed and memorability in solid line.

Password Reuse For the prediction of whether a user choses a new password or reuses an existing one, we see gains in sensitivity up to a true positive rate (TPR) of 65%, trading off a false positive rate (PFR) of up to 18%. After that, increases in sensitivity come at a greater cost in false positive rate.

Memorability For the prediction of whether a user can recall the password, we see sharp gains in sensitivity up to a true positive rate of roughly 85%, trading off a false positive rate of less than 15%. After an FPR of 15%, we do not see a significant increases in TPR for increased FPR.

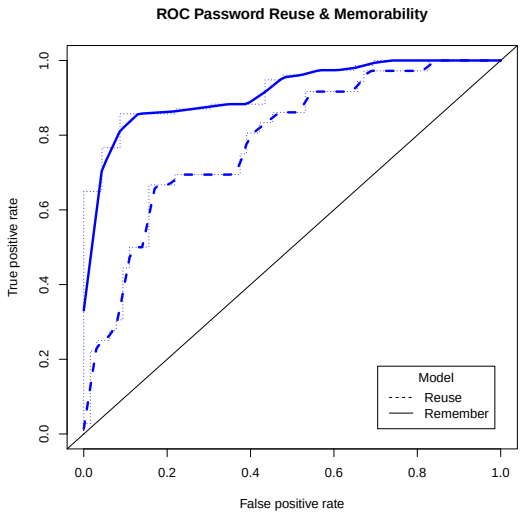


Figure 7: ROC curve for the password reuse and memorability models.

5 Discussion

5.1 Traits impact users' password habits

Agreeable users are more likely to create new passwords; this matches psychological theory because agreeableness is highly correlated with (friendly) compliance in modern facet scales of the Big Five [8]. Hence, the users' habit to reuse passwords reported in earlier studies [13] mediated by users' personality traits.

Will this effect hold up in the real world? First, we need to consider the experimenter expectancy bias [30], that is, the tendency of participants to comply with the experimenter's expectations. Hazelrigg et al. [18] observed that participants with a tendency to comply are prone to be influenced by the experimenter. Hence, one would expect agreeable participants to be more influenced by the experimenter's expectations. However, in early research on the experimenter expectancy bias showed the effect size to be small [3], in general.

Furthermore, we observe that the experimenter's instructions did not actually include a requirement

to create a new password: Participants were only asked to *choose* a password. Consequently, the experiment design made sure the experimenter had no vested interest towards reuse of an existing password or the creation of a new one.

Hence, we believe that the compliance occurring in this case to be related to policies internalized from other contexts. Furthermore, we believe that a tendency of agreeable users to comply to create new passwords in this experiment translates to a tendency to comply to authorities and their endorsed password policies.

5.2 There is a gender gap

Men were four times more likely to create new passwords than women. Whereas men are, thereby, more likely to adhere to the common recommendation to have new passwords, they are exposed to a higher risk to forget these passwords.

Was there a bias inherent in our sample? We sampled international, non-computer-science students eliminating the gender bias prevalent in computer science. From the security awareness questionnaire, we know that there was still a statistically significant difference between genders, with men scoring higher on security awareness than women ($d = 0.75$). We believe that this difference explains part of the different likelihoods to choose either a new password or reuse an old password.

5.3 A new password diminishes recall

Everything else being equal, a user who creates a new password only retains a roughly 50-50 chance to recall the password after a week (cf. Figure 4b on page 12). Surely, in real life users might do a bit better due to stronger incentives to remember the password. However, given the low odds ratio of this predictor, we are certain that this effect persists is real life.

Hence, it is actually a rational strategy for users to reuse existing passwords as this strategy supports memorability considerably.

5.4 Depletion makes things worse

If a user is cognitively depleted at registration time, the likelihood to remember the password shrinks further. Users who are depleted and choose to create a new password are almost certain not to recall it after a week (with a likelihood of remembering of less than 5%).

Given these results, it is clearly counterproductive to create a new password at the end of a long day or after a number of cognitively effortful tasks. In fact, the user is setting himself up for a predictable failure to recall the password again.

5.5 Password strength has little impact

While password strength measured in zxcvbn $\log_{10}$ guesses was not a statistically significant predictor, we observe that the 95% confidence interval bracketed the odds ratio of the password strength at $0.979\ [0.741, 1.293]$. This effect size is very small to near-negligible.

For large differences in zxcvbn $\log_{10}$ guesses, we expect the negative influence of password strength on memorability to be noticeable. For instance, if we compare an exceptionally weak password with $\log_{10}$ guesses of 2 with a strong password with $\log_{10}$ guesses of 12, then we expect the strong password to be less likely to be remembered by a factor of $0.8\ [0.05, 13]$, everything else being equal. However, the negative impact of the password strength is dwarfed by the effect sizes of choosing a new password and of being depleted at registration time.

We evaluated logistic regressions with the password strength metrics CMU PGS, `passwordmeter.com` and the NIST entropy measures as well. All analyses agree to the small effect size of the password strength as predictor of memorability. Thereby, we conclude that the little impact of password strength on recall holds irrespective of the metric evaluated.

5.6 Recommendations

How is it actionable that agreeable users are more likely to choose a new password? One possibility

could be in choice architecture for information security making use of known factors. While security policies today are often written neutrally and a technical language, one could formulate them with an awareness of personality traits and their impacts. While there has been research on how personality traits impact social engineering, such as in the Social Engineering Personality Framework put forward by Uebelacker and Quiel [39], we might as well us a knowledge of personality traits for the good of users and the organizations they belong to. Agreeable users, for instance, might respond more to friendliness in a security policy.

While not statistically significant in this study, extraversion and neuroticism seemed candidates with high enough effect sizes to merit further investigation, which could equally yield suitable soft interventions. At the same time, we caution against discrimination of users, especially when it comes to prejudices with respect to gender.

Choosing a new password unsurprisingly makes recall less likely. Consequently, arbitrary password expiry is putting the user in a bind. Further, we would recommend that not all passwords are created equal. Password policies should not demand from the user to create a new password all the time. It is important that the effort that goes into creating a strong password and into remembering for long-term recall is focused on the passwords that really matter.

Similarly, given that we see only a very small effect size on the impact of the password strength on memorability, we would encourage users to choose long passwords. If a user already chooses a password, it is preferable that the user chooses a good password. "Choose your key passwords like you mean it!"

That cognitive depletion impacts memorability deserves a moment of pause. While earlier research indicated that users create worse passwords under cognitive depletion, we find here that they are also more likely to forget them. Cognitive depletion, thereby, has a double-whammy effect; and registering for an account or creating a new password after a long day's work is a losing strategy. We recommend to create passwords when cognitively replenished.

Let us further consider risk management in fighting security incidents. In such situations, users are bound the a number of cognitively hard decisions and, thereby, prone to getting cognitively depleted. Raising awareness to the resulting risks is important. Equally, being aware that decisions, such as creating passwords, take a hit because of the cognitive depletion, we would recommend to revisit decisions made during the incident response and, for instance, changing the passwords again when well rested.

For all these recommendations, we stress that they are about making dealing with passwords easier for users. If we are burdening users with such an unnatural task, we would be well advised to take it into account how being human plays into the process.

5.7 Ethics

The experiment followed the ethical guidelines of the institution and has received ethical approval. The participants were informed that personal information will be stored in hard and soft copy. The participants were informed of the rough experiment effort and the requirement to come back to the lab in a week, before choosing to participate. The participants were free to participate or not and to withdraw from the experiment at any time. They stated their informed consent in writing.

The participants were paid a compensation of $15 for partial completion and $23 for completing the entire experiment.

The participants data in hard and soft copy was stored securely in an office under lock and key, on stationary machines or laptops with full hard disk encryption. zxcvbn was computed offline. The participants passwords were stripped from username and other PII before being uploaded to CMU's Password Guessability Service (PGS). The data was deleted from CMU's servers after 14 days.

5.8 Ecological Validity

The lab environment was comparable to the lab setting of the ecological validity study of Fahl et

al. [11]. We developed a mockup of GMail, which was visually identical to GMail's account registration page. This was meant to make the task akin real-life settings and to keep the need for the participants to roleplay at a minimum. Even though the experimenter did not disclose that the GMail registration was a mockup, we cannot exclude that participants might have noticed that it was not the real GMail registration page. The experiment included a memorability check for which the participants were asked to return to the lab one week after the registration task. They were to enter the set password in a GMail login mockup. The participants were made aware of this requirement in the initial pre-experiment briefing.

5.9 Limitations

Experiment Design The experiment benefited from random blocked assignment, but was not designed to be double-blind. Care has been taken, however, to design the procedure such that the experimenter would not have a vested interest in the participant's choice to either create a new password or reuse an existing one. Thereby, we sought to stay clear of the experimenter expectancy bias [30].

Lab Sample First, we have chosen to conduct this experiment in the lab and not, for instance, on Mechanical Turk. That has consequences in the sample being recruited largely from international students, even if we focused on a non-computer-science population. Clearly, it impacts external validity and generalizability that we only cover a small age bracket and demographics. At the same time, we note that the Big Five Inventory was first validated on college student samples [22]. Similarly, Baumeister's initial research on the *limited strength model* [4] was on student samples. Hence, internal validity is maintained.

The reason for making this a lab experiment lies in the control of the manipulation. We preferred participants to be observed by an experimenter directly during the battery of cognitively depleting tasks. On the one hand, we found in pretests that that manipulation checks on cognitive effort tasks on MTurk showed significantly lower effect sizes than setups with everything else equal in the lab. Consequently, we expect that MTurk participants would not allow themselves to dip into cognitive depletion as much as participants in the lab. On the other hand, participants who engage in 20-25 minutes of cognitively hard activities, allowing themselves to wear out, deserve having an experimenter present who can offer aftercare.

Sample Size The sample size of $N = 100$ was relatively small for logistic regressions with 4 to 7 predictors, however, not unreasonably so. We put an emphasis on parameter estimation of the odds ratios and their confidence intervals to allow readers to judge effects observed and their magnitude. Further investigation with larger sample sizes is needed, however, to tighten the confidence intervals around the effect sizes.

Strength of Depletion Manipulation The manipulation of the experiment group only yielded 11 participants who reported strong depletion and only 17 participants who were attributed with the effortful condition. The root cause for this is that we used a smaller battery of Stroop tasks compared to similar setups in psychology. Job et al. [20], for instance, use 48 Stroop tasks, where we used 10. As a result, we only have small sample sizes for the effortful and depleted conditions, which leads to larger confidence intervals on those effects.

Lying Participants There is a possibility that participants chose to lie on whether they have re-used an existing password, for instance, to mitigate the risk of a data breach against one of their real e-mail accounts. Lying is considered cognitively effortful. Furthermore, there have been approaches to use cognitive load to amplify the capacity to detect lies [40]. Hence, the induced depletion of the experiment could impact the user's capacity to lie about their password reuse. Then, the depleted condition would see more honest reporting of reuse, the undepleted condition see more lying. We imagine that lying to a simple nominal question, such as "Is this a completely new password?," however, is

still well feasible in any case. Hence, we believe that the impact is small enough that there no significant systematic impact of impaired-capacity-to-lie across conditions.

6 Conclusion

This is the first study to investigate the impact of cognitive depletion on password reuse and memorability. While password habits of users have been investigated in large scale empirical observations [13], we are the first to investigate how personality traits and cognitive depletion impact these habits. Whereas prior work was largely focused on inducing particular styles of password creation [46, 47, 44], we let the user freely choose whether to create a completely new password or whether to reuse an existing one.

We learn that the choice to create of a completely new password or to reuse an existing one is heavily impacted by gender and agreeableness ($OR = 5$). Hence, it is not appropriate to generalize that "the user" is having a particular habit; our reasoning needs to be more refined. We find that men are four times as likely to create new passwords than women. This choice, however, comes at the cost that the likelihood to recall the password after one week is only about 50-50.

The impact of agreeableness is considerable. For each one point increase on a five-point scale the likelihood to choose a new password is five times as high. We also expect that agreeable users are more likely to comply with password policies, in general.

The likelihood to recall the password is impaired by cognitive depletion at registration time. We know now that it hamstrings the user to create a password after a long day or after depleting cognitive tasks. As a consequence, we would recommend to reshape password policies, such that they nudge users to choose passwords when they are fresh.

Finally, we observe that the impact of password strength on memorability is quite limited. Prior research has already found that if users are asked to choose meaningful [31], pronouncable [46] or mnemonic [44] passwords, they can benefit from password strength as well as memorability. We confirmed in this empirical study that users are well able to choose strong passwords, while only suffering from a small reduction in likelihood to remember. Consequently, if a user is already choosing a new password, the user might as well choose a strong one, supported by the strategies proposed.

In conclusion, this research is offering the first differentiated analysis of password habits and memorability, vis à vis of personality traits, current depletion and password strength chosen. While passwords as an authentication method have—rightly—earned a lot of criticism, we face the reality that passwords are still pervasive and often a fallback mechanism for other authentication methods. It will, thereby, remain important to understand the user's actual situation and behavior in face of them.

References

[1] AARTS, H., AND DIJKSTERHUIS, A. How often did i do it? experienced ease of retrieval and frequency estimates of past behavior. *Acta Psychologica 103*, 1 (1999), 77–89.

[2] ADAMS, A., AND SASSE, M. A. Users are not the enemy. *Communications of the ACM 42*, 12 (1999), 40–46.

[3] BARBER, T. X., AND SILVER, M. J. Fact, fiction, and the experimenter bias effect. *Psychological Bulletin 70*, 6p2 (1968), 1.

[4] BAUMEISTER, R., BRATSLAVSKY, E., MURAVEN, E., AND TICE, D. Ego depletion: is the active self a limited resource? *Personality and social psychology 74* (1998), 1252–1265.

[5] BAUMEISTER, R. F., VOHS, K. D., AND TICE, D. M. The strength model of self-control. *Current directions in psychological science 16*, 6 (2007), 351–355.

[6] BONNEAU, J. The science of guessing: analyzing an anonymized corpus of 70 million passwords. In *Security and Privacy (SP), 2012 IEEE Symposium on* (2012), IEEE, pp. 538–552.

[7] BURR, W. E., DODSON, D. F., AND POLK, W. T. Electronic authentication guideline. NIST Special Publication 800-63, NIST, jun 2004.

[8] COSTA, P. T., MCCRAE, R. R., AND DYE, D. A. Facet scales for agreeableness and conscientiousness: A revision of the neo personality inventory. *Personality and individual Differences 12*, 9 (1991), 887–898.

[9] 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.

[10] DELL'AMICO, M., MICHIARDI, P., AND ROUDIER, Y. Password strength: An empirical analysis. In *INFOCOM* (2010), vol. 10, pp. 983–991.

[11] FAHL, S., HARBACH, M., ACAR, Y., AND SMITH, M. On the ecological validity of a password study. In *Proceedings of the Ninth Symposium on Usable Privacy and Security* (2013), ACM, p. 13.

[12] FIELD, A., MILES, J., AND FIELD, Z. *Discovering Statistics Using R*. SAGE Publications, 2012.

[13] FLORENCIO, D., AND HERLEY, C. A large-scale study of web password habits. In *Proceedings of the 16th international conference on World Wide Web* (2007), ACM, pp. 657–666.

[14] FURNELL, S., AND THOMSON, K.-L. Recognising and addressing 'security fatigue'. *Computer Fraud & Security 2009*, 11 (2009), 7–11.

[15] GAW, S., AND FELTEN, E. W. Password management strategies for online accounts. In *Proceedings of the second symposium on Usable privacy and security* (2006), ACM, pp. 44–55.

[16] GOLDBERG, L. R. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology 59*, 6 (1990), 1216.

[17] GROSS, T., COOPAMOOTOO, K., AND AL-JABRI, A. Effect of cognitive depletion on password choice. In *Learning from Authoritative Security Experiment Results (LASER'16)* (July 2016), S. Peisert, Ed.

[18] HAZELRIGG, P. J., COOPER, H., AND STRATHMAN, A. J. Personality moderators of the experimenter expectancy effect: A reexamination of five hypotheses. *Personality and Social Psychology Bulletin 17*, 5 (1991), 569–579.

[19] HOONAKKER, P., BORNOE, N., AND CARAYON, P. Password authentication from a human factors perspective. In *Proc. Human Factors and Ergonomics Society Annual Meeting* (2009), vol. 53, SAGE Publications, pp. 459–463.

[20] JOB, V., DWECK, C. S., AND WALTON, G. M. Ego depletion is it all in your head? implicit theories about willpower affect self-regulation. *Psychological science* (2010).

[21] JOHN, O. P., DONAHUE, E. M., AND KENTLE, R. L. The big five inventory – versions 4a and 54. Tech. rep., Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research, 1991.

[22] JOHN, O. P., NAUMANN, L. P., AND SOTO, C. J. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research 3* (2008), 114–158.

[23] JOHN, O. P., AND SRIVASTAVA, S. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research 2*, 1999 (1999), 102–138.

[24] KELLEY, P. G., KOMANDURI, S., MAZUREK, M. L., SHAY, R., VIDAS, T., BAUER, L., CHRISTIN, N., CRANOR, L. F., AND LOPEZ, J. Guess again (and again

and again): Measuring password strength by simulating password-cracking algorithms. In *Security and Privacy (SP), 2012 IEEE Symposium on* (2012), IEEE, pp. 523–537.

[25] LAURIOLA, M., AND LEVIN, I. P. Personality traits and risky decision-making in a controlled experimental task: An exploratory study. *Personality and Individual Differences 31*, 2 (2001), 215–226.

[26] LAURIOLA, M., RUSSO, P. M., LUCIDI, F., VIOLANI, C., AND LEVIN, I. P. The role of personality in positively and negatively framed risky health decisions. *Personality and individual differences 38*, 1 (2005), 45–59.

[27] MAZUREK, M. L., KOMANDURI, S., VIDAS, T., BAUER, L., CHRISTIN, N., CRANOR, L. F., KELLEY, P. G., SHAY, R., AND UR, B. Measuring password guessability for an entire university. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security* (2013), ACM, pp. 173–186.

[28] MURAVEN, M., TICE, D. M., AND BAUMEISTER, R. F. Self-control as a limited resource: Regulatory depletion patterns. *Journal of personality and social psychology 74*, 3 (1998), 774.

[29] NIELSEN, J. Security and human factors. *Alertbox (November 2000)—http://www. useit. com/alertbox/20001126. html* (2000).

[30] ROSENTHAL, R. Experimenter effects in behavioral research.

[31] SASSE, M. A., BROSTOFF, S., AND WEIRICH, D. Transforming the weakest link: a human/computer interaction approach to usable and effective security. *BT technology journal 19*, 3 (2001), 122–131.

[32] SCHNEIER, B. Secret and lies. *Robert Ipsen* (2000).

[33] SMITH, R. E. The strong password dilemma. *Computer Security Journal 18*, 2 (2002), 31–38.

[34] SPECTOR, Y., AND GINZBERG, J. Passsentence—a new approach to computer code. *Computers & Security 13*, 2 (1994), 145–160.

[35] STANTON, B., THEOFANOS, M. F., PRETTYMAN, S. S., AND FURMAN, S. Security fatigue. *IT Professional 18*, 5 (2016), 26–32.

[36] STROOP, J. R. Studies of interference in serial verbal reactions. *Journal of experimental psychology 18*, 6 (1935), 643.

[37] TICE, D. M., BAUMEISTER, R. F., SHMUELI, D., AND MURAVEN, M. Restoring the self: Positive affect helps improve self-regulation following ego depletion. *Journal of Experimental Social Psychology 43*, 3 (2007), 379–384.

[38] TVERSKY, A., AND KAHNEMAN, D. Availability: A heuristic for judging frequency and probability. *Cognitive psychology 5*, 2 (1973), 207–232.

[39] UEBELACKER, S., AND QUIEL, S. The social engineering personality framework. In *Socio-Technical Aspects in Security and Trust (STAST), 2014 Workshop on* (2014), IEEE, pp. 24–30.

[40] VRIJ, A., FISHER, R., MANN, S., AND LEAL, S. A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling 5*, 1-2 (2008), 39–43.

[41] WÄNKE, M., SCHWARZ, N., AND BLESS, H. The availability heuristic revisited: Experienced ease of retrieval in mundane frequency estimates. *Acta Psychologica 89*, 1 (1995), 83–90.

[42] WEGNER, D. M., SCHNEIDER, D. J., CARTER, S. R., AND WHITE, T. L. Paradoxical effects of thought suppression. *Journal of personality and social psychology 53*, 1 (1987), 5.

[43] WHEELER, D. L. zxcvbn: Low-budget password strength estimation. In *Proc. USENIX Security* (2016).

[44] YAN, J. J., BLACKWELL, A. F., ANDERSON, R. J., AND GRANT, A. Password memorability and security: Empirical results. *IEEE Security & privacy 2*, 5 (2004), 25–31.

[45] ZVIRAN, M., AND HAGA, W. J. Cognitive passwords: The key to easy access control. *Computers & Security 9*, 8 (1990), 723–736.

[46] ZVIRAN, M., AND HAGA, W. J. A comparison of password techniques for multilevel authentication mechanisms. *The Computer Journal 36*, 3 (1993), 227–237.

[47] ZVIRAN, M., AND HAGA, W. J. Password security: an empirical study. *Journal of Management Information Systems 15*, 4 (1999), 161–185.