COMPUTING SCIENCE

Effect of Cognitive Depletion on Password Choice Extended Technical Report

Thomas Groß, Kovila Coopamootoo, Amina Al-Jabri

TECHNICAL REPORT SERIES

No. CS-TR-1496 September 2016

TECHNICAL REPORT SERIES

No. CS-TR-1496 September, 2016

Effect of Cognitive Depletion on Password Choice Extended Technical Report

Thomas Groß Kovila Coopamootoo Amina Al-Jabri

Abstract

Background. The Limited Strength model [3] of cognitive psychology predicts that human capacity to exert cognitive effort is limited and that decision making is impeded once high depletion is reached. Aim. We investigate how password choice differs between depleted and undepleted users. Method. Two groups of 50 subjects each were asked to generate a password. One group was cognitively depleted, the other was not. Password strength was measured and compared across groups. Results. Using a stepwise linear regression we found that password strength is predicted by depletion level, personality traits and mood, with an overall adjusted R 2 = .206. The depletion level was the strongest predictor of password strength (predictor importance .371 and p = .001). Participants with slight effortful exertion created significantly better passwords than the undepleted control group. Participants with high depletion created worse passwords than the control group. Conclusions. That strong depletion diminishes the capacity to choose strong passwords indicates that cognitive effort is necessary for the creation of strong passwords. It is surprising that slight exertion of cognitive effort prior to the password creation leads to stronger passwords. Our findings open up new avenues for usable security research through deliberately eliciting cognitive effort and replenishing after depletion and indicate the potential of investigating personality traits and current mood.

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

Bibliographical details

Effect of Cognitive Depletion on Password Choice
Extended Technical Report

Thomas Groß, Kovila Coopamootoo, Amina Al-Jabri

NEWCASTLE UNIVERSITY
Computing Science. Technical Report Series. CS-TR-1496

Abstract

Background. The Limited Strength model [3] of cognitive psychology predicts that human capacity to exert cognitive effort is limited and that decision making is impeded once high depletion is reached. Aim. We investigate how password choice differs between depleted and undepleted users. Method. Two groups of 50 subjects each were asked to generate a password. One group was cognitively depleted, the other was not. Password strength was measured and compared across groups. Results. Using a stepwise linear regression we found that password strength is predicted by depletion level, personality traits and mood, with an overall adjusted R 2 = .206. The depletion level was the strongest predictor of password strength (predictor importance .371 and p = .001). Participants with slight effortful exertion created significantly better passwords than the undepleted control group. Participants with high depletion created worse passwords than the control group. Conclusions. That strong depletion diminishes the capacity to choose strong passwords indicates that cognitive effort is necessary for the creation of strong passwords. It is surprising that slight exertion of cognitive effort prior to the password creation leads to stronger passwords. Our findings open up new avenues for usable security research through deliberately eliciting cognitive effort and replenishing after depletion and indicate the potential of investigating personality traits and current mood.

About the authors

Dr Thomas Gross is currently a tenured lecturer (assistant professor) in security, privacy and trust at the School of Computing Science at Newcastle University. He is the director of the **Centre for Cybercrime and Computer Security (**CCCS**)**, a UK **Academic Centre of Excellence in Cyber Security Research (ACE-CSR)**. His research interests are in security and privacy as well as applied cryptography and formal methods. He was a tenured research scientist in the Security and Cryptography group of IBM Research - Zurich before that and IBM's Research Relationship Manager for privacy research. Thomas received his M.Sc. (Dipl. Inf.) in Computer Science at the Saarland University, Germany, in 2004. He received his Ph.D. from the Ruhr-University Bochum, Germany, in 2009. His thesis was on the security analysis of standardized identity federation. Thomas is a member of the GI, ACM, IEEE, IACR and EATA, as well as Alumnus of the German National Academic Foundation.

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.

Effect of Cognitive Depletion on Password Choice

Extended Technical Report—Updated 17th January 2019

Thomas Groß
Newcastle University

Kovila Coopamootoo
Newcastle University

Amina Al-Jabri
Newcastle University

Abstract

Background. The Limited Strength model [3] of cognitive psychology predicts that human capacity to exert cognitive effort is limited and that decision making is impeded once high depletion is reached.
Aim. We investigate how password choice differs between depleted and undepleted users.
Method. Two groups of 50 subjects each were asked to generate a password. One group was cognitively depleted, the other was not. Password strength was measured and compared across groups.
Results. Using a stepwise linear regression we found that password strength is predicted by depletion level, personality traits and mood, with an overall adjusted $R^2 = .206$. The depletion level was the strongest predictor of password strength (predictor importance .371 and $p = .001$). Participants with slight effortful exertion created significantly better passwords than the undepleted control group. Participants with high depletion created worse passwords than the control group.
Conclusions. That strong depletion diminishes the capacity to choose strong passwords indicates that cognitive effort is necessary for the creation of strong passwords. It is surprising that slight exertion of cognitive effort prior to the password creation leads to stronger passwords. Our findings open up new avenues for usable security research through deliberately eliciting cognitive effort and replenishing after depletion and indicate the potential of investigating personality traits and current mood.

1 Introduction

Users often set easy-to-remember passwords constructed for example from their wife's name, or recycle and re-use passwords across services [1]. These are predictable and easily guessed. This is because the panoply of separate services mean that users have a list of accounts to manage each with their own login credentials. However, managing and remembering a large number of complex passwords remain a challenge. So far, the question has not been addressed how users create passwords when they are cognitively tired or depleted. In fact, it is an open question whether cognitive effort is *necessary* for the creation of strong passwords.

The limited strength model of cognitive psychology states that human capacity to exert cognitive effort is limited and that decisions as well as effortful tasks are impeded under cognitive depletion [2]. We report on a study with $N = 100$ participants designed to measure the strength of passwords set by cognitively depleted versus non-depleted users. We hypothesise $\mathbf{H}_1$: *Cognitively depleted users create weaker passwords than non-depleted users.*

We replicate existing methods from cognitive psychology [3, 2, 20, 31] to induce cognitive depletion, in particular with thought suppression, impulse control and cognitively effortful tasks. We check depletion manipulation via a Brief Mood Inventory [31]. We measure password strength across groups using a password meter. We evaluate the impact of cognitive depletion on password strength.

Contribution. Our findings indicate that slight exertion of cognitive effort leads to significantly better passwords than an undepleted control group. High depletion leads to worse passwords than an undepleted control group. This is the first study to show the impact of cognitive effort and depletion on password creation. It highlights an important factor for password and usable security research that has not been addressed to date.

2 Background

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

2.1 Strength & Memorability

The use of text usernames and passwords is the cheapest and most commonly used method of computer authentication. The average user has 6.5 passwords, each shared across 3.9 different sites, each user has 25 accounts requiring passwords and type 8 passwords per day [12]. Users have to not only remember the passwords but also the system and userid associated, which password restriction apply to which system and whether they have changed a password and what they have changed it to [1].

Recalling strong passwords is a humanly impossible task since non-meaningful items are inherently difficult to remember [27]. When forced to comply to security policies such as monthly password reset, a large number of users are frustrated [18]. They use strategies such as writing passwords down, incrementing the number in the password at each reset [1], storing passwords in electronic files and reusing or recycling old passwords [18]. While it is possible to create strong and meaningful passwords using pseudo-random combinations of letters, numbers and characters that are meaningful only to the owner [35], four to five passwords are the most a typical user can be expected to use effectively [1].

Thus memory issues impede the strength of password chosen by the user. To help users in setting strong passwords and aid memorability, graphical passwords have been proposed. Methods such as *draw-a-secret* are an improvement on usability of password authentication. Password strength is improved too as even a small subset of graphical passwords constitutes a much larger password space than dictionaries of textual passwords [19]. They work on the principle that humans can remember pictures better than text [29]. However no research has investigated how effortful setting password is for the user nor how depletion states impact password choice and subsequent memorability.

2.2 Cognitive Effort and Depletion

Human beings have a limited store of cognitive energy or capacity [2]. Self-control tasks, choice and decision-making draw from this inner resource. Tasks requiring self-control tasks span across spheres such as controlling attention, emotions, impulses, thoughts and cognitive processing, choice and volition and social processing [3]. In general, all tasks that are cognitively effortful—and thereby *System 2* in the terminology of the dual-process model—draw from the limited cognitive energy. 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 [3] and in unrelated sphere of activity [2]. This phenomenon has been observed in areas of over-eating, -drinking, -spending, under-achievement, and sexuality [3].

An underlying question of this research is whether the creation of strong passwords is a cognitively effortful task. If that is the case, then we expect to observe that the creation of passwords is impaired under cognitive depletion. On such an observation, we can further conclude that cognitive effort is necessary for password creation. Given that cognitive depletion permeates different activities and is yielded by a variety of self-control tasks, we can therefore expect that password creation will be impaired by the user's other effortful activities.

2.2.1 Beliefs

In this context, it is an important question whether all people are equally cognitively depleted. Interestingly, a person's beliefs have an influence on the level of that person's cognitive depletion. There is a line of research in (motivational) psychology investigating the impact of beliefs on the nature of human attributes. A classical example is the belief whether intelligence is fixed or malleable [5, 10]. It turns out that implicit beliefs about willpower as a limited resource [20] impact the extent of cognitive depletion. Consequently, individuals who believe in unlimited willpower are less affected by cognitive depletion. This effect impacts our experiment because participants are not equally affected by the manipulation inducing cognitive depletion, and we expect to see differing depletion levels in the experiment group.

2.2.2 Personality Traits

While beliefs or mindsets of persons constitute personality traits already, we expect other personality traits to influence the capacity to bear cognitive effort as well as the strength of chosen passwords. Capacity theories of self-control conceptualise it as a dispositional trait like construct that differ across individuals. Thus people high in dispositional self-control will have more resources at their disposal than individuals lower in trait self-control. In addition, certain people are dispositionally motivated to act in a certain way such as over-eating, -drinking. Personality traits has already been linked with security research, for example impulsive individuals are more likely to fall for phishing e-mails while trait-based susceptibility to social engineering attacks is recognised [32].

2.2.3 Affect

Security tasks including password security often leads to user frustration [18]. While affect states impact decisions, they also influence cognition [28]. Affect states enable recall of mood congruent information that might influence judgments, or the heuristic adopted to make decisions [6].

In addition the active regulation of emotion or mood deplete self-control resources and invoke ego depletion [2]. Regulating affect often requires the individual to overcome the innate tendency to display emotions while negative affect, induced by demanding and frustrating tasks, is implicated in development of ego depletion.

3 Method

3.1 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 from Newcastle and Northumbria University, of mainly international background (common countries included Oman, China and Iraq). 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 $p = 1.00$.

3.2 Procedure

The experiment was designed to enable a comparison of the influence of cognitive depletion on password strength. 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 flavour.

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. Figure 1 depicts the experiment design.

3.2.1 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.

3.2.2 Inducing Cognitive Depletion

We induce cognitive depletion for the experiment condition, reproducing manipulation components of Baumeister et al. [31]. 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.

As discussed in Section 2.2.1, we expect participants in the experiment condition to be affected by the induction of cognitive effort to differing degrees. Especially the implicit theories about willpower have been shown to have a significant effect on cognitive depletion [20]. Consequently, we will control the strength of the manipulation with a manipulation check based on a brief mood inventory (Section 3.3.2) evaluated in the Results Section 4.1.

1. Thought supression task. In the experiment condition, the participants are shown a lot white bear and asked *not* to think of the white bear, a procedure following Wegner et al. [33]. 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, but not instructed suppress it. The control condition, is not cognitively depleting as the participants do not need to suppress their thoughts.

2. Impulse control task. This task is adapted from Muraven et al. [26]. 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 to cross out all letters 'e'. This rule does not require impulse control and is thereby non-depleting.

3. Cognitively effortful task. We used the Stroop task [30] 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.

3.3 Measures

3.3.1 Password Strength

We tested multiple password meters, such as the Microsoft password meter and finally settled for the password meter Web site[1] because it uses an interval scale and makes the components of the password score transparent. Each component, such as 'number of characters' or presence of 'numbers', gives a bonus or malus for the overall score. All

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

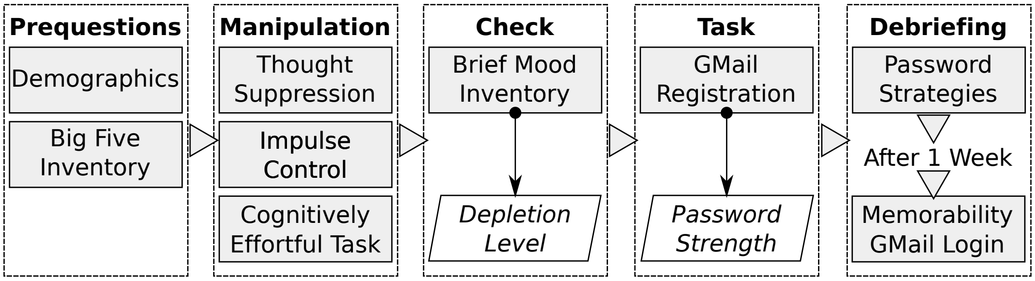

Figure 1: Overview of the experiment procedure. The control group did manipulation tasks with similar structure and flavor, yet without the depleting condition.

component scores were recorded individually and their sum computed as password score. Whereas the password meter itself caps the scores at 0 and 100, our final score could be negative or greater than 100.

The password meter does not account for weaknesses such as the use of dictionary words or personal identifiable information (e.g., name, username) as part of the password. By the NIST password guidance [7], those conditions, especially failing the dictionary test, make weak passwords. Hence, we adjusted the obtained password scores with penalties if the password contained:

- an unmodified dictionary word (-25),
- part of the user's real name (-50),
- the username (-50), or
- the user's student id (-50).

The dictionary words we checked were the large list of the Openwall wordlist collection[2], intended primarily for use with password crackers such as John the Ripper and with password recovery utilities. For the username, we argue that this information is often at the disposal of an adversary in offline attacks. For instance, for the Linux /etc/passwd file, the username is the first field of each entry, the real name in often encoded in the comment field. We obtain a final password strength score on an interval scale, with values between -100 and 150. The password strength obtained from this procedure was nearly normally distributed across the participants.

In addition to the password meter score, we evaluated the NIST password entropy according to the heuristic given in the NIST Special Publication 800-63 [7] and submitted the passwords to the CMU Password Guessability Service (PGS) [24], however both methods offered limited differentiation across the range of password strengths.

3.3.2 Brief Mood Inventory

Earlier research found that cognitive depletion can be checked with a brief mood inventory, either the Brief Mood Introspection Scale (BMIS) [25, 2] or a short form. We use a short form of a brief mood inventory used as manipulation check in Baumeister's research [31], including the dimensions (a) excited, (b) thoughtful, (c) tired, (d) happy, (e) worn out, (f) sad, (g) angry, (h) calm, rated on 5-point Likert-type items between 1 Disagree strongly and 5 Agree strongly, with 3 Neither agree nor disagree as central point. Baumeister el al. [31] found that tiredness and feeling worn out are significantly affected by cognitive depletion and can therefore be used as self-report manipulation check.

3.3.3 Big Five Inventory

The personality traits of the users were queried with a 60-item Berkeley Big Five Inventory (BFI) [14, 21, 22]. The inventory measures the traits (a) Openness to experience, (b) Conscientiousness, (c) Ex-

[2] http://www.openwall.com/wordlists/

traversion, (d) Agreeableness, and (e) Neuroticism, with a 5-point Likert-type items between 1 Disagree strongly and 5 Agree strongly computing the scores as means of items for each domain.

4 Results

All inferential statistics are computed with two-tailed tests and at an alpha level of .05.

4.1 Manipulation Check

We used the brief mood inventory introduced in Section 3.3.2 as manipulation check on the cognitive depletion, following a methodology of Baumeister et al. [31].

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 $p = .000 < .05$; worn out: $U = 669$, $Z = -4.145$, significance $p = .000 < .05$). As expected following Baumeister et al. [31] 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 $r = 0.63$ and for feeling being worn out is $r = 0.42$. That constitutes a large effect on feeling tired and a medium to large effect on feeling worn out. Consequently, we reject the null hypothesis that the control and experiment group are equally depleted across conditions. This suggests that the cognitive depletion of the participants has been induced by the manipulation.

4.2 Password Strength Score

The distribution of the Passwordmeter password strength score is measured on interval level and is not significantly different from a normal distribution, Saphiro-Wilk, $W(100) = 1$, $p = .652$. The distribution of the Password Guessability Service (PGS) results are measured on interval level and significantly different from a normal distribution, Saphiro-Wilk, $W(100) = .0.921$, $p < .001$. We continue the analysis with the Passwordmeter password strength score. Figure 2 given an overview of the visual inspection of the password strength scores. It contains a histogram in comparison with the normal distribution, the probability density estimate and the QQ-plot between the sample data and the theoretical distribution. We computed Levene's test for the homogeneity of variances. For the password meter scores, the variances were not significantly unequal (a) for experiment and control condition, $F(1,98) = 3.369$, $p = .069 > .05$, and for (b) for gender, $F(1,98) = 1.378$, $p = .243 > .05$.

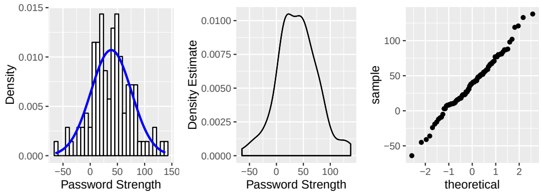

Figure 2: Visual inspection of the distribution assumptions on the sampled password strength scores.

Univariate Analysis of Variance (GLM). We conducted a Univariate Analysis of Variance (GLM) with Type III Sums of Squares—robust against unequal sample sizes—with password strength as dependent variable. We used condition, gender and the Brief Mood Inventory (BMI) items as fixed variables, the Big Five Inventory (BFI) as covariates.

(a) There was a significant effect of gender, $F(1,60) = 6.824$, $p = .011$, partial $\eta^2 = .102$. (b) We observed a significant effect of BMI_Tired, $F(4,60) = 3.687$, $p = .009$, partial $\eta^2 = .197$. (c) BMI_Calm had a significant effect, $F(4,60) = 4.264$, $p = .004$, partial $\eta^2 = .221$. Other factors did not show significant effects. The corrected model offered a variance explained of $R^2 = .533$ (adjusted $R^2 = .229$).

4.3 Automated Linear Regression.

The impact on the password strength score was analyzed with a multi-predictor forward stepwise linear regression. The linear regression has an adjusted $R^2 = .206$. The studentized residual was close to a normal distribution. The outcomes of the gender, Big Five (5) and brief mood inventory (8) were predictors on the password strength score as target variable. The gender did not have a significant effect in the linear regression. We provide the effects, effect sizes (η^2 and ω^2) and coefficients of the linear regression in Tables 5 and 6 in the Appendix. Section 4.4 contains the posthoc power analysis for this regression. We give details of the automated data preparation of the regression first and will subsequently describe the effects in decreasing order of predictor importance.

Depletion Level from Brief Mood Inventory The SPSS automated data preparation of the linear regression merged categories of BMI_Tired to maximise the association with the target. Table 4 in the Appendix contains an exact overview of the SPSS automated data preparation. We accept this grouping and name the cases introduced by SPSS and call it the *depletion level* of (a) non-depleted, (b) effortful, and (c) depleted. Strongly disagree, disagree slightly and neither agree nor disagree were grouped as BMI_Tired_T $= 0$. We label this case as depletion level non-depleted. The agree slightly of BMI_Tired was transformed into BMI_Tired_T $= 1$, which we label as depletion level effortful. BMI_Tired of Agree strongly was transformed into BMI_Tired_T $= 2$, which we label as depletion level depleted.

We evaluated the derived depletion level as variable for further analysis. First, we evaluated Levene's test for the homogeneity of variances. For the password meter scores, the variances were not significantly unequal for depletion level, $F(2,97) = 1.997$, $p = .141 > .05$.

A One-way ANOVA was computed with the password strength score as dependent variable. There was a significant effect of the depletion level on the password strength, $F(2,97) = 5.449$, $p = .006 < .05$. We observe a medium to large effect size: $\eta^2 = 0.10$. As the ANOVA was computed with unequal sample sizes, we employed Scheffé and Games-Howell as robust post-hoc tests. Scheffé uses the harmonic mean sample size 18.335. Both Scheffé and Games-Howell reported the depleted case significantly different from the effortful case, Scheffé $p = .006 < .05$ and Games-Howell $p = .036 < .05$. Neither of the tests found a significant difference between the non-depleted and the other two cases.

Accepting the grouping of the SPSS automated data preparation, we have: Of the control group, 49 participants were consequently 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. Table 1 contains an overview of descriptive statistics over these groups.

On this grouping, we observe an odds ratio that a participant is fully depleted of

$$OR_{\text{depleted}} = 12.25 \qquad 95\% \text{ CI } [1.5, 99.8].$$

The odds ratio be being somewhat affected by depletion (i.e., effortful and depleted combined) is:

$$OR_{\neg\text{non-depleted}} = 57.52 \qquad 95\% \text{ CI } [7.36, 449.75].$$

Effects of Cognitive Depletion. The depletion level was indeed the most important predictor in the regression (significance $p = .001 < .05$, predictor importance= .371). The depletion level was recoded to make the non-depleted condition the baseline.[3] The effortful level, that is only slightly depleted, had a coefficient of 19.027 (significance $p = .044 < .05$). The depleted level had a coefficient of -31.623 (significance $p = .006 < .05$). We observe a medium effect size of the depletion level, $\omega^2 = .097$. Consequently, we reject the null hypothesis.

We summarize the descriptive statistics of password strength score by depletion level in Table 1 and depict them in Figure 3.

The descriptive statistics on password guessability determined by PGS show the same overall outcome in terms of means of password guesses as well as percentage of passwords determined as unguessable (cf. Table 2). The proportional difference of PGS unguessable passwords across depletion level

[3] The LASER paper [15] reported coefficients with the depleted condition as baseline. Otherwise, both analyzes are equivalent.

was significant (Fischer Exact Test, two-tailed, $p = .033 < .05$).

We evaluate the effect size of the number of unguessable passwords with odds ratios. Comparing the cases effortful and depleted pairwise against non-depleted we have the following odds ratios of number of unguessable passwords:

$$OR_{\mathsf{effortful}} = 3.62 \qquad 95\% \text{ CI } [1.18, 11.06]$$
$$OR_{\mathsf{depleted}} = 0.41 \qquad 95\% \text{ CI } [0.05, 3.45].$$

Effects of Mood. BMI Thoughtfulness and Calmness had significant effects. Strong disagreement to thoughtfulness implied stronger passwords (significance $p = .018 < .05$, predictor importance = .251, coefficient 40.072). Strong disagreement to calmness implied stronger passwords (significance $p = .012 < .05$, predictor importance = .172, coefficient 38.799). Both BMI Thoughtfulness and Calmness constitute small effects, at $\omega^2 = .052$ and $\omega^2 = .044$.

Effects of Personality Traits. Of the Big Five personality traits, the BFI Agreeableness score was the most important predictor on the password strength (significance $p = .025 < .05$, predictor importance = .137, coefficient 14.649), where higher agreeableness significantly implied stronger passwords. The BFI Extraversion was a notable yet non-significant negative predictor on password strength (significance $p = .108 > .05$, predictor importance = .069, coefficient -11.538). Both BFI Agreeableness and Extraversion constitute small effects, at $\omega^2 = .034$ and $\omega^2 = .013$.

4.4 Power Analysis

All power analyses are computed with G*Power 3.1 [11], aiming at a significance level of $\alpha = .05$.

A Priori Power. We conducted an a priori power analysis for F-Tests with Omnibus One-Way ANOVA. We intended detect at least a medium effect size $f = 0.25$ of cognitive depletion on password strength with a power of 80%. This scenario implies a sample size on the order of 128 participants.

We have elected to use a constrained random assignment with balanced gender and balanced time of day for the lab appointments to control variability introduced by those factors.

5 Discussion

This study applied the methodology of previous cognitive depletion studies [3, 2, 20, 31] to a ubiquitous security context. We observe that cognitive effort is a major predictor of password strength. Moderate cognitive exertion leads to stronger passwords than in an non-depleted and in depleted states. Strong depletion leads to weaker passwords than in moderate exertion and non-depleted states.

This is in accord with Kahneman's observation that initial effortful activity introduces a bias towards exerting further cognitive effort [23]. This outcome can also be explained with Selye's arousal curve [9], an inverse U-shaped relation between the activity of the stress system and the quality of a human's performance, yielding an optimum performance under moderate stress. This result vouches for further investigation, in particular to what extent this observation can be operationalized to improve the quality of password choice.

Our analysis also showed the impact of mood and personality, hence indicates the importance of studying other human dimensions. The results on the brief mood inventory are surprising in themselves, in particular because participants who reported themselves as not thoughtful or not calm chose better passwords. This result can substantiate the explanation of Selye's arousal curve as a possible explanation. In any case, these results ask for the investigation of the influence of current stress and mood on password choice, in particular whether negative emotions such as fear are involved.

For personality traits, it is plausible that BFI Agreeableness has an impact on the password strength, because the NEO PI classifies Compliance one of its facets, and hence postulates a tendency to avoid conflict and cooperate. Therefore, we assume users with high agreeableness to comply to password policies, as well. However, the results on the BFI ask for further investigation, as the experiment

Table 1: Descriptive statistics of password strength via password meter by condition and depletion level.

Condition	Depletion Level	N	Mean	Std. Dev.	Std. Error	Min	Max
Control	Non-depleted	49	40.65	30.97	4.43	-45	121
	Depleted	1	16.00	-	-	16	16
	Total	50	40.16	30.87	4.37	-45	121
Experiment	Non-depleted	23	33.30	33.81	7.05	-19	102
	Effortful	17	57.12	45.07	10.93	-24	138
	Depleted	10	11.10	45.97	14.54	-64	70
	Total	50	36.96	42.99	6.08	-64	138
Total		100	38.56	37.27	3.73	-64	138

Table 2: Descriptive statistics of password guessability determined by the CMU Password Guessability Service (PGS) by condition and depletion level. PGS declares passwords "unguessable" at 2.E+13 guessing attempts.

Condition	Depletion Level	N	Mean	Std. Dev.	Std. Error	Unguessable Pwds % (#)
Control	Non-depleted	49	4.85E+12	7.19E+12	1.03E+12	22.4% (11)
	Depleted	1	40945471	-	-	0% (0)
	Total	50	4.75E+12	7.15E+12	1.01E+12	22% (11)
Experiment	Non-depleted	23	2.37E+12	5.69E+12	1.19E+12	13%(3)
	Effortful	17	7.83E+12	8.49E+12	2.06E+12	47.1% (8)
	Depleted	10	1.71E+12	5.22E+12	1.65E+12	9.1% (1)
	Total	50	4.09E+12	7.11E+12	1.01E+12	24% (12)
Total		100	4.42E+12	7.10E+11	7.10+11	23% (23)

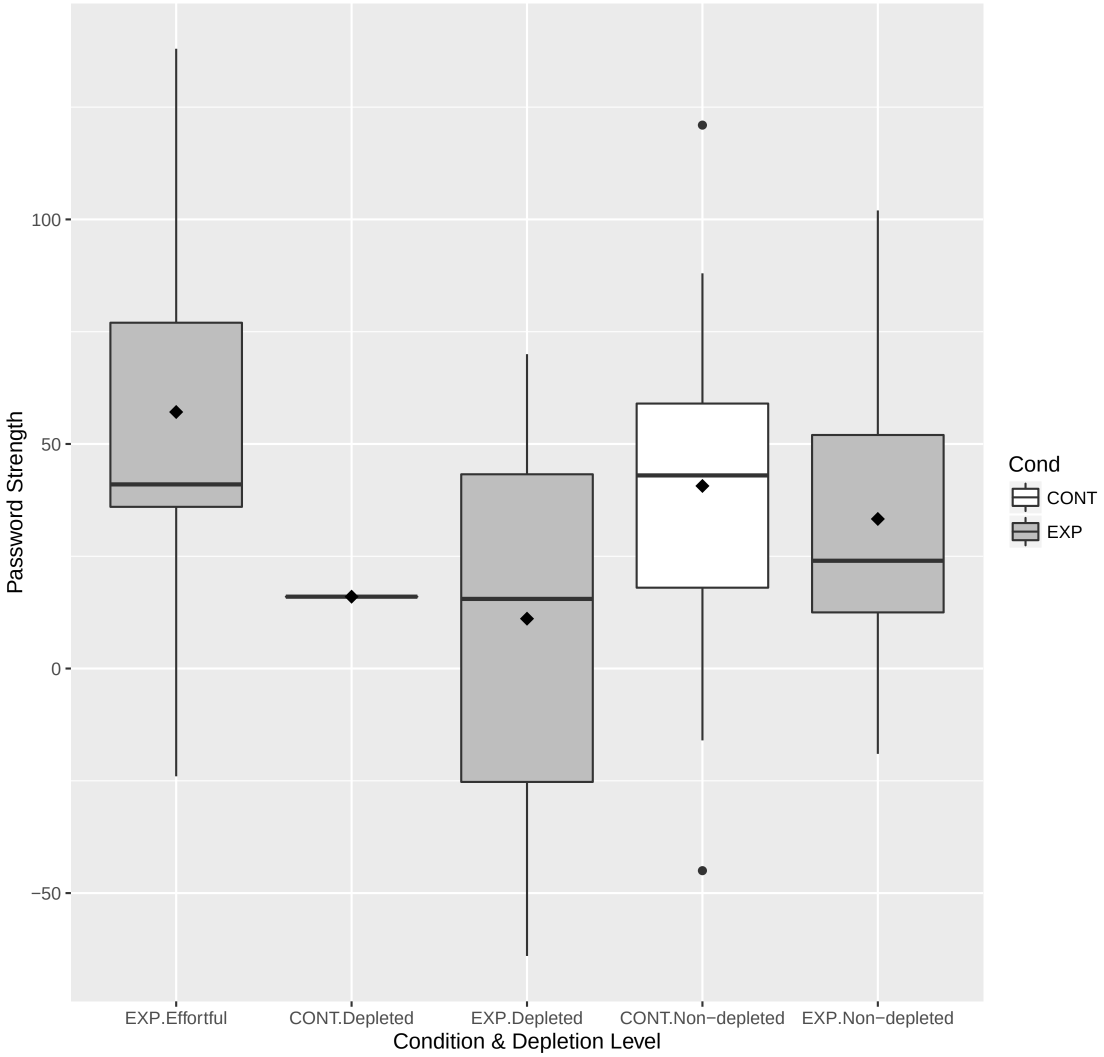

Figure 3: Boxplots of the password strength score by depletion level and experiment condition. The diamond shows the group mean. 49 participants of the control condition were non-depleted, one participant was depleted. 23 participants of the experiment condition were non-depleted, 17 classified as effortful, 10 as depleted.

cannot distinguish whether the participants sought to please the experimenter, constituting a confounding variable, or whether the effect of compliance persists in real-world scenarios. It is notable that BFI Conscientiousness, the tendency to show self-discipline and be dutiful, did not have a significant effect on the password strength.

5.1 Ethics

The experiment followed the ethical guidelines of the university and has received ethical approval. The participants were informed that personal identifiable data will be stored in hard and soft copy and have consented to the experiment procedures. 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 paid a time compensation of $15 for partial completion and $23 for completing all components of the 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. 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.

5.2 Ecological Validity

We developed a mockup of GMail, which was visually identical to GMail's account registration page. In this sense the experiment is generalisable to real-life settings. 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.

5.3 Limitations

We account for limitations of the given experiment and offer mitigation options for future experiments where appropriate.

Experiment Design. Whereas the experiment was balanced in that the participants of experiment and control group did manipulation tasks of similar structure that only differed in the depletion condition, the experiment was not designed to be double-blind. The experimenter knew which condition the participants were in. Future experiments can use a second experimenter for the tasks after the manipulation unaware of the depletion state.

Strength of Manipulation. Even if the effect size of the manipulation in odds ratio was noteworthy at $OR_{\mathsf{depleted}} = 12.25$, the absolute number of participants that reported strong depletion was low ($n = 11$). Future experiments will need to achieve stronger depletion throughout and manipulate a slight cognitive effort stimulus deliberately.

The cognitive depletion manipulation was only partially successful with 28 participants out of 50 reporting slight or strong agreement with tiredness. The effect size for at least some reported depletion in odds ratio was $OR_{\neg\mathsf{non-depleted}} = 57.52$. Earlier studies, such as [20] used 48 Stroop task items, while our study only contained 10 items. Future experiments can increase the cognitive effort by increasing the number Stroop task items.

Unequal Sample Sizes. Due to the differing impact of the manipulation on the participants, the grouping by depletion level yielded unequal sample sizes. This could have confounded analysis with One-way ANOVA. Consequently, we have employed an Univariate Analysis of Variance with Type III Sums of Squares robust in this situation. Alternative approaches with random re-sampling to groups with equal sample sizes agreed to the analysis outcomes.

Low granularity on depletion levels. Depletion levels have only been differentiated on a 5-point Likert-type scale. Further categorisation of these levels will be beneficial to investigate when cognitive effort promotes or inhibits security behaviour. Future experiments can mitigate this limitation with

Table 3: Correlation of password strength metrics.

	Passwordmeter	Zxcvbn	PGS
Passwordmeter		.43	.46
Zxcvbn	.43		.58
PGS	.46	.58	

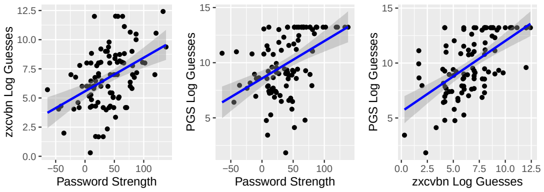

Figure 4: Scatterplots with regression lines and 95% CI for different password metrics.

either a 9-point Likert-type scale or a Visual Analogue Scale (VAS).

Future experiments could be further improved with an invasive post-task cognitive depletion manipulation check, such as a long Stroop Task counting the number of errors. Placed after the experiment task, additional depletion inflicted by the Stroop Task would not contaminate the measurement in the control group.

Password measures. The password strength measurements considered in this study all come with weaknesses: The password meter has limitations in being purely heuristic. The other two measures lack in differentiation across the range of password strengths. The NIST entropy estimate only offers a low differentiation between password strengths. The results of the CMU Password Guessability Service (PGS) are not normally distributed and come with a static cut-off at which the service considers a password "unguessable", conflating the strengths of secure passwords to a single value.

As a further password strength measurement tool, we considered Zxcvbn, which also measures the likely number of guesses an adversary will need to crack the password. We have compared the three tools passwordmeter.net (NIST-corrected score), PGS and Zxcvbn (both log10 number of guesses) and found that their results are only to some extent pair-wise correlated. A correlation matrix with Pearson's correlation coefficients showed statistically significant results. Table 3 shows the correlation matrix between the different password strength metrics tested, while Figure 4 contains the scatter plots with regression lines.

Newly created vs. reused password. Not all participants created a new password. However, we asked participants in the post-study questionnaire whether they created a new password or whether they reused an existing password. There was a noticeable difference across depletion levels, where the effortful condition had the highest proportion of reused password. However, the difference in proportions across depletion levels was not statistically significant, $p = .215$, Fischer Exact Test.

Low Adjusted R^2 in Linear Regression. The adjusted $R^2 = .206$, hence the automated linear regression accounts for 20.6% of the variability, adjusted for the number of predictors in the model. We observe that the variability in the experiment group as well as in the participants with a depletion level of effortful or depleted was higher than in the control group.

6 Conclusion

We offer the first comprehensive study of cognitive effort and depletion in a security context. We conclude that cognitive effort is a *necessary* condition for the creation of strong passwords, which in turn implies an involvement of *System 2* in terms of the dual-process model. It is an intriguing observation that slight cognitive effort improves password strength and that cognitive depletion diminishes password strength. It has far-reaching consequences for the design of password user interfaces and password policies. First, we observe that the user's cognitive effort and depletion may be more important than solely concentrating on password complexity requirements. Second, the user's cognitive depletion may yield an alternative explanation for and substantiate earlier research on the security compliance budget [4]. Third, our investigations indicate practical amendments to password policies ("Only set a new password when you feel fresh and awake.") and possible HCI interventions to strengthen password behavior (e.g., by inducing cognitive effort before the password generation).

Acknowledgments

We are grateful to Roy Maxion and Pam Briggs for the insightful discussions on this research. We are grateful for the insightful comments of the LASER reviewers, the LASER'16 workshop participants and of Chris Kanich. We thank Lorrie Cranor, Sean Segreti and Blase Ur for the support in the use of the CMU Password Guessability Service (PGS). This work was supported by the EU FP7 Project FutureID (GA n^{o}318424) and the EPSRC Research Institute in Science of Cyber Security (RISCS; EP/K006568/1 ChAISe). We would like to thank Ingolf Becker for the suggestion at LASER'16 that a post-task manipulation check could well be cognitively effortful without confounding the password choice task. We thank Malte Elson for the discussion on the limited-strength model and the 2016 preregistered multilab replication attempts of the cognitive depletion effect.

References

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

[2] 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.

[3] 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.

[4] BEAUTEMENT, A., SASSE, M. A., AND WONHAM, M. The compliance budget: managing security behaviour in organisations. In *Proceedings of the 2008 workshop on New security paradigms* (2009), ACM, pp. 47–58.

[5] BLACKWELL, L. S., TRZESNIEWSKI, K. H., AND DWECK, C. S. Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child development 78*, 1 (2007), 246–263.

[6] BOWER, G. H. Mood congruity of social judgments. *Emotion and social judgments* (1991), 31–53.

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

[8] CARTER, E. C., AND MCCULLOUGH, M. E. Publication bias and the limited strength model of self-control: has the evidence for ego depletion been overestimated? *Frontiers in psychology 5* (2014), 823.

[9] CHROUSOS, G. P. Stressors, stress, and neuroendocrine integration of the adaptive response: the 1997 hans selye memorial lecture. *Annals of the New York Academy of Sciences 851*, 1 (1998), 311–335.

[10] DWECK, C. S. *Self-theories: Their role in motivation, personality, and development.* Psychology Press, 2000.

[11] 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.

[12] 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.

[13] FORDYCE, T., GREEN, S., AND GROSS, T. Investigation of the effect of fear and stress on password choice. In *Proceedings of the 7th Workshop on Socio-Technical Aspects in Security and Trust* (2018), ACM, pp. 3–15.

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

[15] 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.

[16] HAGGER, M. S., CHATZISARANTIS, N. L., ALBERTS, H., ANGGONO, C. O., BATAILLER, C., BIRT, A. R., BRAND, R., BRANDT, M. J., BREWER, G., BRUYNEEL, S., ET AL. A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science 11*, 4 (2016), 546–573.

[17] HAGGER, M. S., WOOD, C., STIFF, C., AND CHATZISARANTIS, N. L. Ego depletion and the strength model of self-control: a meta-analysis. *Psychological bulletin 136*, 4 (2010), 495.

[18] 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.

[19] JERMYN, I., MAYER, A. J., MONROSE, F., REITER, M. K., RUBIN, A. D., ET AL. The design and analysis of graphical passwords. In *Usenix Security* (1999).

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

[23] KAHNEMAN, D. *Thinking fast and slow*. Farrar, Strauss, 2011.

[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] MAYER, J. D., AND GASCHKE, Y. N. The experience and meta-experience of mood. *Journal of personality and social psychology 55*, 1 (1988), 102.

[26] 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.

[27] 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.

[28] SCHWARZ, N., AND CLORE, G. L. How do i feel about it? the informative function of affective states. *Affect, cognition, and social behavior* (1988), 44–62.

[29] SHEPARD, R. N. Recognition memory for words, sentences, and pictures. *Journal of verbal Learning and verbal Behavior 6*, 1 (1967), 156–163.

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

[31] 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.

[32] 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.

[33] 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.

[34] WHEELER, D. L. zxcvbn: Low-budget password strength estimation. In *USENIX Security Symposium* (2016), pp. 157–173.

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

A Corrections and Changes

A.1 Correction on Saphiro-Wilk on Password Strength [2019/01/16]

In Section 4.2 “Password Strength Score” of the published LASER’16 paper [15], we reported for the password meter “$D(100) = .99$, $p = .652$” mentioning it being greater than the significance level as well as for PGS log10 guesses “$D(100) = .59$,” reporting a triple-zero p-value.

Those tests were misreported and flagged in an analysis with the R package statcheck after the camera-ready submission.

We recomputed them on the original dataset with the following outcome: For the password meter we have Shapiro-Wilk $W(100) = 1$, $p = .652$; for PGS log10 guesses we have Saphiro-Wilk, $W(100) = .0.921$, $p < .001$. These corrections do not change the conclusions drawn from the tests.

No further statcheck errors were found.

A.2 Power Observed [2019/01/16]

An earlier version of extended technical report contained a post-hoc power and positive predictive value (PPV) consideration. We removed the post-hoc discussion as we could not reliably quantify power loss due to, e.g., unequal group sizes, and retained the results of the original *a priori* power analysis as reported.

B Thoughts on Replication Attempts

We invite and aim at supporting replication attempts of this work. We offer some thoughts in hindsight, how replications could be informed.

Evidence on the Limited Strength Model. It is beneficial to consider the evidence available in psychology to this date. In 2010, Hagger et al. [17] conducted a systematic meta analysis of 83 studies testing the effect of ego depletion. As overall effect of cognitive depletion, the study found an averaged corrected standardized mean difference of

Table 4: Automated data preparation of the Automated Forward Stepwise Linear Regression: The automated data preparation has merged categories to maximize association with the target password strength for BMI_Thoughtful, BMI_Tired, BMI_Wornout, BMI_Calm.

Transformed Field	Label	Included Original Categories
BMI_Tired_T $= 0$	Non-depleted	Disagree strongly, disagree a little, neither agree nor disagree
BMI_Tired_T $= 1$	Effortful	Agree a little
BMI_Tired_T $= 2$	Depleted	Agree strongly
BMI_Thoughtful_T $= 0$		Disagree strongly
BMI_Thoughtful_T $= 1$		Disagree a little
BMI_Thoughtful_T $= 2$		Neither agree nor disagree, agree a little
BMI_Thoughtful_T $= 3$		Agree strongly
BMI_Calm_T $= 0$		Disagree strongly
BMI_Calm_T $= 1$		Disagree a little, neither agree nor disagree, agree a little, agree strongly

Table 5: Effects of the Automated Forward Stepwise Linear Regression.

Source	Sum of Squares	df	Mean Square	F	Sig.	Imp.	η^2	ω^2
Corrected Model	37,143.052	8	4,642.882	4.210	$< .001$			
BMI_Tired_T	15,580.722	2	7,790.361	7.065	.001	.371	.113	.097
BMI_Thoughtful_T	10,516.094	3	3,505.365	3.179	.028	.251	.076	.052
BMI_Calm_T	7,197.450	1	7,197.450	6.527	.012	.172	.052	.044
BFI_Agreeableness	5,749.695	1	5,749.695	5.214	.025	.137	.042	.034
BFI_Extraversion	2,897.897	1	2,897.897	2.628	.108	.069	.021	.013
Residual	100,349.588	91	1,102.743					
Corrected Total	137,492.640	99						

Note: $R^2 = .27$, adjusted $R^2 = 0.206$, Akaike Information Criterion Corrected = 711.125.

Table 6: Coefficients of the Automated Forward Stepwise Linear Regression. The coefficients BMI_Tired_T $= 2$, BMI_Thoughtful_T $= 3$ and BMI_Calm $= 1$ have been set to zero because they are redundant. Note that the depletion level has been recoded in this technical report to make the non-depleted condition the baseline. Consequently, the effects reported here differ from the short LASER paper [15].

Model Term	Label	Coef.	Std. Err.	t	Sig.	95% Conf. Interval		Imp.
						Lower	Upper	
Intercept		3.559	35.561	0.1	.920	-67.078	74.197	
BMI_Tired_T $= 1$	effortful	19.027	9.318	2.024	.044	0.517	37.536	.371
BMI_Tired_T $= 2$	depleted	-31.623	11.331	-2.791	.006	-54.132	-9.115	.371
BMI_Thoughtful_T $= 0$		40.072	16.704	2.399	.018	6.892	73.252	.251
BMI_Thoughtful_T $= 2$		19.405	8.514	2.279	.025	2.493	36.318	.251
BMI_Calm_T $= 0$		38.799	15.187	2.555	.012	8.632	68.967	.172
BFI_Agreeableness		14.649	6.415	2.283	.025	1.906	27.392	.137
BFI_Extraversion		-11.538	7.117	-1.621	.108	-25.675	2.600	.069

$d^{+} = 0.62$, 95% CI $[0.57, 0.67]$. The reported effect sizes are aligned with the assumptions we have made in our *a priori* power analysis.

However, Carter and McCullough [8] indicated in 2014 that the field of ego depletion research shows strong signals of publication bias. The paper notes the presence of small-study effects and the danger of overestimating effect sizes as a result. The authors report a conspicuous lack of statistically non-significant findings.

Finally, we consider the 2016 multilab preregistered replication attempt of Baumeister's ego depletion work [16].[4] The replication attempt reported a summary effect of Cohen's $d = 0.04$, 95% CI $[-0.07, 0.15]$. Given that our LASER'16 study sought to apply instruments used in Baumeister's work, e.g., [31], the low effect size observed in a large-scale replication ($N = 2,141$) asks us to be prudent. This prudence includes not only cognitive depletion as a phenomenon, but also the manipulation and manipulation-check instruments used.

Manipulation Instruments. Hagger et al. [16] included a systematic meta analysis of manipulation instruments used in its appendix, overlapping with the instruments used in our study. Hence, expected effect sizes for manipulation checks can be well-informed from their work.

Manipulation Check Instruments. We would find it advisable to use well-established and vetted manipulation check for cognitive depletion, where we have used Visual Analogue Scale (VAS) in this context after the LASER'16 publication. The Brief Mood Inventory (BMI) variant we used could well be replaced with a measurement instrument with fidelity to our research intention, while obtaining a more accurate manipulation check.

One subtlety to consider here is that measures of cognitive depletion are often also inducing cognitive effort. For that reason, Baumeister's proposal of a Brief Mood Inventory as a proxy for cognitive depletion holds some appeal as pre-task manipulation check. However, a cognitively effortful measurement instrument, such as counting errors on a defined battery of Stroop tasks (cf., Job et al [20]), could be employed *after* the password choice task. In that case, the cognitive effort induced by the measurement would not confound the task any longer, while offering a measurement how much cognitive depletion was still present at the end of the password choice task.

Password Strength Measures. Finally, in subsequent studies on the effects of stress and fear on password choice [13], we considered other outcome measures. We concluded that the password meter metric used in this study and the Password Guessability Services (PGS) make it more challenging to replicate the study, because we have no control over the current version available to future replications. Consequently, we decided for zxcvbn [34] log10 guesses as metric, yielding the advantage that we could store the exact zxcvbn version used in the study and that the dictionaries used by zxcvbn could be externalized. While log10 guesses measures a different construct than a linear password meter heuristic score, we found in our studies that the resulting variables are well correlated.

Power. The effect size observed in the large-scale replication is clearly lower than the one we assumed in our *a priori* power analysis (corresponding to a medium-effect two-tailed sensitivity (Cohen's $d = 0.566$) at 80% power under equal-sized and normally-distributed groups). Given the effect size estimates observed in the recent replications, we would aim at a study powered for a small effect size.

Analysis Methods. We note that the automated data preparation of the SPSS Automated Forward Stepwise Linear Regression introduced a grouping used in our analysis. We would find it a sound change to introduce three conditions in the manipulation (undepleted, effortful, and depleted) *a priori* through well-defined and controlled levels of exposure to cognitively effortful tasks, instead.

[4] Replication OSF Repository: `https://osf.io/jymhe/`