Investigation of the Effect of Fear and Stress on Password Choice

Tom Fordyce
Newcastle University
Newcastle upon Tyne, UK

Sam Green
Newcastle University
Newcastle upon Tyne, UK

Thomas Groß
Newcastle University
Newcastle upon Tyne, UK

ABSTRACT

Background. The current cognitive state, such as cognitive effort and depletion [22], incidental affect or stress may impact the strength of a chosen password unconsciously.

Aim. We investigate the effect of incidental fear and stress on the measured strength of a chosen password.

Method. We conducted two experiments with within-subject designs measuring the zxcvbn [55] log10 number of guesses as strength of chosen passwords as dependent variable. In both experiments, participants were signed up to a site holding their personal data and, for the second run a day later, asked under a security incident pretext to change their password. (a) **Fear.** $N_F = 34$ participants were exposed to standardized fear and happiness stimulus videos in random order. (b) **Stress.** $N_S = 50$ participants were either exposed to a battery of standard stress tasks or left in a control condition in random order. The zxcvbn password strength was compared across conditions.

Results. We did not observe a statistically significant difference in mean zxcvbn password strengths on fear (Hedges' $g_{av} = -0.11$, 95% CI $[-0.45, 0.23]$) or stress (and control group, Hedges' $g_{av} = 0.01$, 95% CI $[-0.31, 0.33]$). However, we found a statistically significant cross-over interaction of stress and TLX mental demand.

Conclusions. While having observed negligible main effect size estimates for incidental fear and stress, we offer evidence towards the interaction between stress and cognitive effort that vouches for further investigation.

CCS CONCEPTS

• **Security and privacy** → **Authentication**; **Human and societal aspects of security and privacy**; • **Human-centered computing** → Empirical studies in HCI;

KEYWORDS

password choice, cognitive effort, depletion, stress, fear

ACM Reference format:
Tom Fordyce, Sam Green, and Thomas Groß. 2017. Investigation of the Effect of Fear and Stress on Password Choice. In *Proceedings of 7th ACM Workshop on Socio-Technical Aspects in Security and Trust, Orlando, Florida, USA, December 2017 (STAST'2017),* 13 pages.
https://doi.org/10.475/123_4

STAST'2017, December 2017, Orlando, Florida, USA

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

1 INTRODUCTION

While recently heralded as its days being numbered [6, 7], username and password are still the predominant authentication mechanism. At the same time, password choice makes for an archetypical security task, in which users are focused on reaching a primary goal (accessing a Web site), while the security of the password is only a secondary goal. Hence, results on users' password choice can be informative beyond the act of choosing a password itself.

While there is a body of research on user habits in password choice [15], recent research also aimed at investigating how the user's current cognitive or affective state impacts the strength of chosen passwords. Imagine a user being depleted after a long day's work or being stressed out after a painful social interaction.

Given that such cognitive and affective states are found to impact *inter alia* executive function, working and declarative memory, effects on security decision making are plausible. This may hold for for security specialists fighting a security incident, experiencing stress and cognitive depletion in the process, as much as for off-the-street users, experiencing stress and cognitive depletion in everyday life. The current state of the user could then unconsciously impair the strength of the password choice.

While Florêncio et al. [16] already considered password strategies for finite-effort users, Groß et al. [22] investigated the effect of cognitive effort and depletion [3, 28] on password choice in depth, concluding that cognitive effort is a necessary condition for strong passwords. By their account, cognitive depletion would impair password choice significantly.

We investigate the *effect of fear and stress on password choice* as an archetypical security task. We focus on incidental stress, that is, stress that is unrelated to the security task at hand.

It is rather elusive to induce stress independently from cognitive effort, because the host of stress induction instruments cause require the participant to exert cognitive effort at the same time. The reason for that is that by Baumeister's *limited strength model* [3] also governs the use of willpower. Stress induction techniques, however, no matter whether they are cognitive [34], physical [36], or social [31] require the participant to exert willpower as well to keep going, to persevere in the task.

For that reason, we have not only designed an experiment that induces stress, but also a second experiment that induces fear. We have chosen to induce fear with a standardized video stimulus, which does not yield any cognitive effort whatsoever.

In this paper, we contribute two studies establishing the effect of incidental fear as well as of stress on password choice. To our knowledge, these are the first studies that induce fear and stress in a password choice scenario. In addition, we investigate the interaction between stress and cognitive effort. We offer research synthesis between these two studies and the earlier work on cognitive effort in a network meta analysis to put these results in context of one another.

2 PRELIMINARIES

This study is founded on the principles of parameter and interval estimation [12], that is, we are less interested in null-hypothesis significance testing (NHST). Instead, we seek to quantify the magnitude of effects and to offer confidence intervals to bracket the likely effect sizes in the population.

While NHST has received its share of criticism and observed fallacies [19, 39], we aim at gaining robustness through complementing NHST with estimation methods [12].

2.1 Effect Size and Interval Estimation

We use effect sizes for correlated (within-subjects) means with Hedges' g_{av}. Therein, we follow recommendations of Cumming [12] and Lakens [32]. We use an interval estimation for correlated samples by Algina and Keselman [2, 12].

The extended technical report for this paper [18] contains the exact formula and reasoning for these methods.

2.2 Network Meta Analysis

Meta analysis refers to a set of statistical techniques to combine the results of multiple studies [11, 12]. *Network meta analysis* (NMA), in particular, specializes on combining results of studies with different treatments in the same meta analysis simultaneously. For instance, in this paper we consider studies which all studied the impact of certain "treatments," such as stress, fear or cognitive effort, on password strength. Schwarzer et al. [48] offer an introduction to the method itself, while Binod et al. [38] offer a good overview of such techniques available in R. In this work, we use the R package netmeta, which implements frequentist and graph-theoretical techniques following Rücker [44].

3 BACKGROUND

3.1 Password Choice

In terms of overall user password behavior, the average user is said to have 6.5 passwords, each shared across 3.9 different sites, each user has 25 accounts requiring passwords and type 8 passwords per day [15].

It has been argued that the recall in password authentication itself is a humanly impossibly task, because non-meaningful items are inherently difficult to remember [47]. This is aggravated by typical password policies asking users to comply to arcane procedures, such as monthly password reset. Such policies cause users to feel frustrated and security-fatigued [27]. Consequently, users are naturally employing alternative 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 [27].

Password reuse, in particular, having been observed as overall tendency and individual preference, has received further attention in recent research [13].

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 [57], from qualitative research it was found that four to five passwords are the most a typical user can be expected to use effectively [1].

3.1.1 Password Strength Measurement. There are controversies around how to measure password strength soundly and reliably [5], where earlier methods such as the NIST password entropy heuristic (NIST Special Publication 800-63 [9]) have received criticism.

A recent research direction is to practicably estimate the number of guesses needed by an adversary to break the password [29, 55].

In this work, we considered the Password Guessability Service (PGS) [29] hosted by CMU as well as the Dropbox password meter zxcvbn [55], which both output an estimate of the number of guesses. We chose zxcvbn as final measurement instrument.

3.2 Affect

Affect is the experience of an emotion or feeling, where we also consider the observable affect, that is, behavior serving as indicator of affect.

While there exist a number of conceptualizations for affect, mood and emotions as well as a body of mature research in psychology [14, 33], we consider the Russel's core affect conceptualization [46] as a guiding work. Russel considers the dimensions activation-deactivation and pleasure-displeasure as foundational.

We make a distinction between two kinds of affect [41]:

- *integral affect* (also task-related affect) refers to the experienced feelings with respect to a stimulus, and
- *incidental affect* refers to feelings such as mood states that are independent of a stimulus but can be misattributed to it or can influence decision processes.

For our enquiry into fear and stress, Russel's core affect model is especially interesting as it allows us to classify distress and fear in the common quadrant of displeasure/activation.

3.3 Fear and Fear Appeals

Affects have been considered in security mostly concerning fear, especially in relation to fear appeals. *Fear appeals* [17, 42, 56] are messages designed to motivate a certain behavior by eliciting fear.

While fear appeals are are operating with integral/task-related fear, we consider incidental fear and stress in this work.

Boss et al. [8] gave an overview of fear appeals research in information security, pointing out that most prior studies in the space did not measure fear directly. In general, Ruiter et al. [45] pointed out that coping information is bound to be more important in yielding a protection motivation than risk warnings and fear arousal.

3.3.1 Affect Elicitation. There has been considerable research on affect elicitation. For instance, the first ten chapters of the *Handbook of Emotion Elicitation and Assessment* [10] are concerned with different elicitation methods. Not all elicitation methods are created equal, though. Westermann et al. found considerable differences in effectiveness and validity of mood induction procedures (MIP) [54]. In this study, we focus on Film MIP without instructions which was a category recommended by Westermann et al. From inspecting this past research, we conclude that the chosen MIPs are sound to elicit affect in participants.

Affect elicitation with stimulus films has received a systematic treatment in affective psychology [20], where our stimulus videos are drawn from Rottenberg et al.'s work [43]. Here, we find particular recommendations for video segments validated for designated target emotions, such as fear or happiness.

We observe that it was reported among others by Rottenberg et al. [43, pp. 103] that fear is a challenging emotion to elicit discretely. In their experiments with stimulus films, such as *The Silence of the Lambs*, they found that the films also elicited tension and interest in equal measures, however they argued that the confluence of fear, interest and tension is indeed natural.

We note that affect elicitation with stimuli films has been validated in within-subject experiment designs [43]. The variable "within-subject design" was taken into account in a comprehensive meta analysis on MIPs [54], where the impact of this variable was not found to be statistically significant ($r = .08$ for films/story MIPs.).

3.3.2 Affect Measurement. While Coan and Allen [10] give an overview of a range of affect measurement methods, we focus on self-report instruments such as the Positive and Negative Affect Schedule (PANAS-X) [53]. The questionnaire primarily measures the valence of reported affect (positive or negative) and its content (e.g., fear or joviality). The questionnaire was designed to reliably measure affect while still being easily administered. The full 60-item schedule usually takes participants about 10 minutes.

3.4 Stress

Selye [49] originally defined stress as the *the nonspecific response of the body to any demand made upon it*, and distinguished between eustress and distress, hyperstress and hypostress.

It was postulated that some levels of stress may improve performance and that performance will deteriorate once an optimal range of arousal is passed. With respect to arousal in habit formation, such a relationship was also formulated as the Yerkes-Dodson law [51].

3.4.1 Stress Elicitation. Stress can be elicited in experiments predominantly by cognitive, physical or social instruments. Liao and Carey [34] offer an overview of lab stress elicitation methods.

Existing experimental protocols often combine validated cognitive, physical and social stress tasks, where such tasks have been used to induce psychological stress and receive a cardiovascular response [34]. In addition, there exist instruments, such as *Trier Social Stress Test* (TSST) [31], which are elaborate protocols to induce stress in multiple stages.

3.4.2 Stress Measurement. While stress can be measured psychophysiologically from heart rate variability and skin conductance, a number of instruments have been proposed to measure stress in self-report questionnaires.

Partially, researchers used affective instruments, such as the State-Trait Anxiety Inventory (STAI) [50] to gauge stress.

Partially, researchers developed and used specialized stress scales, such as the Dundee Stress State Questionnaire (DSSQ) [35] or its short form, the Short Stress State Questionnaire (SSSQ) [25, 26].

3.5 Cognitive Effort

Research in cognitive effort has a long history in psychology, for instance starting from Kahneman's work on effort and attention [28].

Baumeister et al. [3] first proposed that human beings have a limited store of cognitive energy, formulated in the *limited strength model of cognitive effort*. Willpower and self-control are said to draw from this inner resource as well as cognitively hard tasks. Examples include controlling attention, emotions, impulses, thoughts and cognitive processing, choice and volition and social processing [4]. In general, all tasks that are cognitively effortful draw from the limited cognitive energy.

3.5.1 Cognitive Effort Measurement. Cognitive effort as a construct can, for instance, be measured by tasks that require cognitive effort themselves. In those examples the error rate on the tasks will indicate cognitive depletion. These measurement methods come at the disadvantage that they change the participants' state by inducing cognitive depletion themselves.

Baumeister et al. consequently proposed to use a Brief Mood Introspection Scale (BMIS) [3, 37] or a short form [52] as a proxy to measure cognitive effort through items such as "being tired" and "being worn-out."

Other research focused on the measurement of the cognitive effort needed for a task, the *task load*. A notable measurement instrument along this lines that stood the test of time is the NASA Task Load Index (TLX).

4 AIMS

RQ 1 (Study 1: Fear). *To what extent do elicited affects happiness and fear impact password strength?*

Table 1 gives an overview of the operationalization of this research question. As independent variable (IV), we have elicited affect with the two levels: Fear and Happiness.

We intend to check that the manipulation was successful by evaluating the Positive and Negative Affect Schedule (PANAS-X) [53] on fear and joviality. The null hypothesis of this manipulation check is $H_{mc,F,0}$: *There is no mean difference between either fear or joviality between conditions.* We call the manipulation successful if this null hypothesis is rejected.

The null hypothesis of the overall experiment is $H_{F,0}$: *There is no mean difference between zxcvbn* log10 *guesses between conditions.* The corresponding alternative hypothesis $H_{F,1}$ is *The zxcvbn* log10 *guesses differ between conditions.*

RQ 2 (Study 2: Stress). *To what extent does elicited stress impact password strength?*

Table 2 operationalizes this research question. As independent variable (IV), we have elicited stress with the two levels: Stress and Control.

We check the success of the manipulation with two kinds of instruments: the Short State Stress Questionnaire (SSSQ) [26], considering overall stress and distress, as well as the State-Trait Anxiety Inventory (STAI) [50], considering state_anxiety.

The null hypothesis of this manipulation check is $H_{mc,S,0}$: *There is no mean difference between either stress, distress, or state_anxiety.* We call the manipulation successful if this null hypothesis is rejected.

The null hypothesis of the overall experiment is $H_{S,0}$: *There is no mean difference between zxcvbn* log10 *guesses between stressed and control condition.* The corresponding alternative hypothesis $H_{S,1}$ is *The zxcvbn* log10 *guesses differ between conditions.*

Table 1: Operationalization of Study 1: Effect of Fear on Password Strength.

	Levels	Instrument	Intervention/Variable
IV: Affect	Fear Happiness	Stimulus Video [43]	*Silence of the Lambs* *When Harry Met Sally*
IV Check	Fear Happiness	PANAS-X [53]	fear joviality
DV: Pwd Strength		zxcvbn [55]	log10 Guesses

Table 2: Operationalization of Study 2: Effect of Stress on Password Strength.

	Levels	Instrument	Intervention/Variable
IV: Stress	Stress	Serial Subtraction Task [34]	cont. sub. 7 from 9095, 1.5 min cont. sub. 13 from 5245, 1.5 min
		Isometric Handgrip Task [36]	30% max strength, 2.5 min
		Social Stress akin to TSST [31]	results judged; fail: start over
	Control	Balanced: Serial addition	cont. add 7 to 9095, 1.5 min cont. add. 13 to 5245, 1.5 min
		Balanced: Isometric Handgrip	30% max strength, till discomfort
IV Check	Total Stress Distress Anxiety	SSSQ [25, 26, 35] STAI [50]	stress distress state_anxiety
Cognitive Effort	Task Load Mental Demand	TLX [23, 24]	tlx tlx_mental
DV: Pwd Strength		zxcvbn [55]	log10 Guesses

5 METHOD

For reproducibility and scientific integrity, the study has been registered at the Open Science Framework (OSF)[1]. The comprehensive report [21] of all analyses computed is registered at the OSF, as well. Analyses, graphs and statistical reporting in this paper were computed directly from the data using the R package knitr.

5.1 Common Approach

5.1.1 Within-Subjects Lab Experiment. Both Studies were conducted simultaneously as within-subject experiments with two conditions, that is, each participant goes through both conditions. We determined in a constrained random block assignment in which order each participant was exposed to the conditions, while maintaining balanced sub-sample sizes.

We stress that affect elicitation and mood induction procedures have been experimentally and meta-analytically validated for within-subject settings [43, 54]. To ensure that a MIP stimulus of a preceding session does not confound a subsequent session, we leave a break of 24 hours between sessions.

We have further chosen to run the studies as a lab experiment and not as a study on Amazon Mechanical Turk (AMT), primarily because the stress manipulations required physical presence of the participants. Given that we induced stress as well as fear, we deemed it ethical that the experiments were conducted by a physically-present experimenter, who could offer information about the experiment in person, allow participants to withdraw from the experiment in dignity and ensure that participants are not leaving the experiment overly disturbed.

[1] osf.io/3cd9h

5.1.2 Sampling. The participants for both studies were recruited independently from one another.

We have chosen to run the experiments as within-subject design with the given sample sizes to gain at least 80% power for medium effect sizes, after an *a priori* power analysis. A two-tailed dependent-samples t-test requires a sample size of $N = 34$ to reach 80% power at a significance level of $\alpha = .05$. The effect size estimates were chosen to be smaller than observed medium-large effects reported by Groß et al. [22].

In both studies, the participants were largely recruited from university staff and students.

5.1.3 Overall Procedure. In both studies, participants are asked to register on a Web site, which stores personal information as well as sensitive data about them, such as personality traits and psychometric test results. The participants are made aware of the sensitivity of the data. There was no password policy imposed on the participants. Participants were asked to "choose" a password, that is, they were not asked to refrain from reusing prior passwords.

The experiments are conducted over two days, to let the effects of prior manipulations subside. The break between both runs was at least 24 hours, but no more than five days. All participants returned for the second run. No participant withdrew from the study.

When the participants returned for the second day, they are informed that they need to set a new password for their personal data under the pretext of a security incident. The system enforced that they could not repeat exactly the same password.

For both experiments, we first elicited an affective state, then asked the participants to register an account with a password, and finally had a post-task manipulation check to evaluate how well the affect elicitation worked. We note here that the manipulation check

was deliberately placed *after* the password task, to ensure that the affect during the task was at least as strong as the one measured.

The participants were only told in the debriefing of each study that the experiment's true purpose was about password strength. Study 2 included a debriefing questionnaire on the modalities of password choice made.

5.1.4 Password Strength Measurement. Password strength was measured as log10 number of password guesses as evaluated with an offline zxcvbn [55] with standard dictionaries.

5.2 Ethics

Both studies followed the institution's ethics guidelines and were approved in its ethics process.

Affect Elicitation. Participants were exposed to mild discomfort in the form of stress or fear, yet not more so than in daily life. The stimuli used have been validated in affective psychology and stress research and been found appropriate for the use in experiments with adult participants.

The experiments were conducted in a face-to-face setting to ensure the experimenter could offer aftercare should the participants feel uncomfortable or upset after a session.

Informed Consent and Opt-Out. Participants were informed of the requirements (two lab sessions) of the studies in advance.

Participants received a consent form, could ask questions before and during the experiments, and were informed that they could withdraw from the experiment at any time. All participants were able to exercise informed consent.

Deception. The participants were deceived in that we did not disclose that our main interest was the password choice. Instead, the personality traits, affect and stress measurements were presented as part of a personality profile Web site.

Participants received a debriefing in which the true purpose of the experiment was explained.

Compensation. Participants were reimbursed for their time spent in the experiment at the institution's customary rate for lab experiments. We set the policy that participants would be reimbursed even if they chose to withdraw from the study.

Data Protection. We ensured data protection and privacy of the participants' sensitive information. Records were anonymized and stored on an encrypted hard disk. The zxcvbn metrics on the participants' passwords were computed offline.

5.3 Study 1: Effect of Fear

5.3.1 Participants. The participants were recruited through e-mailing lists with in the university on students in computer science, mathematics, and statistics as well as flyers. The sample consisted of 25 students and 9 participants from a range of professions, incl. nursing, teaching, and management.

A total sample of $N_F = 34$ was recruited, 11 women and 23 men. We note that for mood induction procedures, gender has not been found a statistically significant confounding variable [54].

88% of the participants were Caucasian, 9% from Pacific or Asian islands, one of mixed Asian/Caucasian ethnicity. The majority of participants had a BSc degree (79%), 12% a high school degree, 3% an MSc and 9% a PhD. Table 3 shows the distribution of gender and age ($M = 29.29$, $SD = 14.82$).

Table 3: Demographics of Study 1: Fear

Gender	
Female	32%
Male	68%

Age	
18–29	79%
30–44	3%
45–59	9%
60+	9%

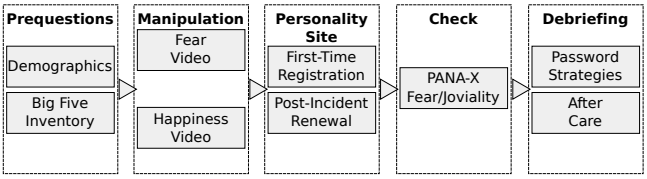


Figure 1: Experiment design of Study 1: Fear.

The sample size of the experiment was determined *a priori* with a sensitivity to detect medium differences between dependent means of Cohen's $d = 0.5$ at 80% power.

5.3.2 Procedure. We offer an overview of the exact procedure in Figure 1. Participants were constrained randomly assigned to be exposed to either the fear or the happiness stimulus in the first session.

5.3.3 Manipulation. Participants were asked to watch standardized stimulus videos that induce either happiness or fear [43] from the *Handbook of emotion elicitation and assessment* [10]. To elicit fear, we selected a specified scene from the *Silence of the Lambs*, in which an FBI agent is trying to find a psychopath, the movie's villain, in a dark basement.

To elicit happiness, we selected a specified scene from *When Harry Met Sally*, in which the two main characters are sitting in a café talking about a fake orgasm and Sally starts to fake an orgasm to prove her point.

The effectiveness of both stimuli has been documented in the corresponding research in affective psychology [43].

5.3.4 Manipulation Check. The manipulation check allowed us to test whether the manipulation was successful as well as to compute a correlation between the measured strength of the affect and the password strength.

As post-task manipulation check, we administered the *Positive and Negative Affect Schedule* (PANAS-X) [53] as a self-report questionnaire to evaluate the participants' current affects.

The PANAS-X is scale is based on 5-point Likert-items. anchored on 1 – "very slightly or not at all," 2 – "a little," 3 – "moderately," 4 – "quite a bit," and 5 – "extremely."

We restricted PANAS-X to the items pertaining to the variables na (negative affect), pa (positive affect), fear and joviality, which yields a 40-item questionnaire. We anchored PANAS-X on "at the present moment."

Table 4: Demographics of Study 2: Stress

Gender	
Female	52%
Male	48%

Age	
18–29	100%
30–44	0%
45–59	0%
60+	0%

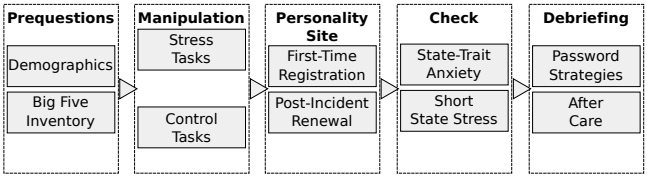


Figure 2: Experiment design of Study 1: Stress.

5.4 Study 2: Effect of Stress

5.4.1 Participants. The $N_S = 50$ participants were recruited from university students, mostly with computer science background. Table 4 includes their gender and age distribution ($M = 20.92$, $SD = 1.05$).

In terms of gender, the sample was was roughly balanced, with 26 women and 24 men. We observe that gender is relevant for stress induction procedures. Research of, e.g., Kelly et al. [30] showed that "women are more likely than men to report higher levels of negative affect and fear to social stress challenges," while they "do not significantly differ in the extent of autonomic arousal and cortisol reactivity."

The sample size for the stress experiment was determined *a priori* with a sensitivity to detect differences between means of dependent groups of Cohen's $d = 0.4$ with 80% power. This target effect size was chosen to be smaller than the effects observed by Groß et al. [22] to analyze for incidental stress as alternative explanation.

5.4.2 Procedure. Figure 2 depicts the experiment design and procedure for the second study on stress. Participants were constrained randomly assigned to be exposed to the stress condition either in the first or the second session.

5.4.3 Manipulation. The manipulation in the experiment condition consisted of two stressful tasks combined with induction of social stress. In the control condition, the participants completed balanced tasks which did not induce stress.

Serial Subtraction Task. We induced cognitive stress by the *serial subtraction task*, one of the most used tasks to induce psychological stress and receive a cardiovascular response [34]. In the experiment condition, the participant is asked to continually subtract an one- or two-digit prime number from a four digit number. The participants completed two serial subtraction tasks with 7 and 13 counting down from 9095 and 5245, respectively. Each task lasted 1.5 min. When a participant miscomputed a value, the participant was asked to start over. This was framed as a test of cognitive ability.

In the control condition, participants were asked to continuously add 7 or 13 to 9095 and 5245, respectively, for 1.5 min. If they made a mistake, they were told the correct result, but not to start over.

Isometric Handgrip Task. The *isometric handgrip task* has been used to induce physical stress in terms of cardiovascular response [36]. An electronic hand dynamometer was used to measure the participant's maximal grip strength.

In the experiment condition, the participants were asked to hold the grip at least at 30% of their maximal grip strength for 2.5 min. Should they go under 30%, the experimenter would sternly tell them to keep it above "Keep it above [30% of their max]." The experimenter took notes of the participants's performance.

In the control condition, the participants were asked to hold the grip at 30% max strength till they start to feel uncomfortable and, then, to stop. No notes were taken in the control condition.

Social Stress. Part of the protocol was to induce social stress in the experiment condition akin to the methods of the *Trier Social Stress Test* (TSST) [31]. While we did not replicate the TSST exactly, we took inspiration from it. In the serial subtraction task, participants were told that their results would be reviewed with the principal investigator of the study. TSST also used serial subtraction, and similar to the TSST, participants were asked to start over, when they made a mistake.

In the isometric handgrip task's experiment condition, the experimenter was standing behind the sitting participants. The experimenter issued stern warning and made notes whenever the participant slipped under 30% of the max grip strength.

5.4.4 Manipulation Check. We used two instruments to check the stress of participants: the Short State Stress Questionnaire [25, 26] and the State-Trait Anxiety Inventory [50].

Short State Stress Questionnaire. The Short State Stress Questionnaire (SSSQ) [26] is self-report questionnaire on state stress, an abridged version of the Dundee State Stress Questionnaire (DSSQ) [35]. The short version contains 24 questions, formalized as 5-point Likert items anchored on 1 – "Strongly Agree," 2 – "Agree," 3 – "Neither Agree nor Disagree," 4 – "Disagree," and 5 – "Strongly Disagree." Questions referred to the time "in this moment."

The SSSQ contains three factors: engagement, distress, and worry. We considered the overall stress, the sum of these three, as well as distress.

State-Trait Anxiety Inventory. The State-Trait Anxiety Inventory for Adults (STAI-AD) [50] is a 40-question self-report questionnaire. We are interested in the temporary construct of state anxiety, that is, "how you feel right now." It uses 4-point Likert items anchored on 1 – "Not At All," 2 – "Somewhat," 3 – "Moderately So," and 4 – "Very Much So."

NASA Task Load Index. The NASA Task Load Index (TLX) [23, 24] is a standardized and validated instrument to measure the overall task load (strongly related to cognitive effort). It includes subscales for mental, physical and temporal demand as well as performance, effort and frustration level. It measures these sub-scales on visual analogue scales (VAS), which we projected on the interval $[-10, +10]$. For the overall TLX measurement, the different sub-scales are weighted by the subjective workload ranks.

5.4.5 Debriefing Questionnaire. The debriefing of Study 2 inquired on multiple aspects of the participants' password choice for both conditions. The questions included:

- reuse: Has the password or a variant been used previously?
- frequency: How often has the password been used in the past?
- last_use: When has the password been last used?
- strategy: Why has the password been chosen that way?

6 RESULTS

As a general rule, statistics were computed with a significance level of $\alpha = .05$. We used two-tailed dependent-samples tests throughout. We consider the manipulation checks of each study as a test family and report p-values with Bonferroni-Holm multiple-comparisons corrections as $p_{\mathrm{MC}(n)}$, where n is the number of comparisons made. We do not correct the password strength comparison or the order-effect analysis to ward against Type-II errors.

6.1 Study 1: Effect of Fear

6.1.1 Descriptives. The study measured PANAS-X fear and joviality as manipulation checks and zxcvbn log10 guesses as dependent variable. Table 5 offers an overview of the descriptive statistics of the fear experiment ($N_{\mathrm{F}} = 34$).

Table 5: Descriptive statistics of Study 1: Fear

(a) Elicited Affect: Fear

	PANAS-X		zxcvbn
	fear	joviality	log10 Guesses
M	2.91	2.35	6.39
SD	0.68	0.88	2.89

(b) Elicited Affect: Happiness

	PANAS-X		zxcvbn
	fear	joviality	log10 Guesses
M	1.12	3.47	6.74
SD	0.17	0.83	3.28

6.1.2 Manipulation Check: PANAS-X. As manipulation check, we compared PANAS-X measurements on fear and joviality across the two conditions Fear and Happiness. Figure 3 compares the distributions of the treatments for each measurement.

Assumptions. We analyzed the difference of both conditions for outliers based on the Outlier Labeling Rule. We further checked for outliers with the Mahalanobis Distance D^2 and, finally, concluded that no outlier needed to be capped or removed.

We tested the distribution of differences between conditions for normality with Shapiro-Wilk. While the differences of the joviality measurements were sufficiently normally distributed ($W = 0.96, p = .199$), we found that the differences of the fear measurements were not normally distributed, $W = 0.94, p = .049$.

While the dependent-samples t-test is deemed to some extent robust against violations of normality, we complement it with a Wilcoxon Signed-Rank test.

Success of the Fear/Happiness Manipulations. Comparing across conditions, the mean fear was statistically significantly greater under elicited fear than under elicited happiness, $t(33) = 15.79, p_{\mathrm{MC}(2)} < .001$, Hedges' $g_{\mathrm{av}} = 4.15$, 95% CI [2.64, 4.61]. We observed a very large effect.

The Wilcoxon signed-rank test confirmed this result, $V = 595, p < .001$.

Comparing across conditions, the mean joviality was statistically significantly less under elicited fear than under elicited happiness, $t(33) = 15.79, p_{\mathrm{MC}(2)} < .001$, Hedges' $g_{\mathrm{av}} = 1.38$, 95% CI [0.91, 1.9] This was a large effect.

We rejected the null hypothesis $\mathrm{H}_{\mathrm{mc,F},0}$. Consequently, the manipulation check showed that the stimulus videos *The Silence of the Lambs* and *When Harry met Sally* indeed caused fear and happiness in the participants.

6.1.3 Password Strength. We compared the password strength measured in zxcvbn log10 guesses across conditions.

Assumptions. By the Outlier Labeling Rule and the evaluation of Mahalanobis distance D^2, there were no significant outliers.

Differences Between Conditions. The log10 number of guesses is not statistically significant across conditions, $t(33) = -0.65, p = .520$, Hedges' $g_{\mathrm{av}} = -0.11$, 95% CI [−0.45, 0.23]. The effect size was trivial. We failed to reject the null hypothesis $\mathrm{H}_{\mathrm{F},0}$.

The log10 guesses are statistically significantly correlated across conditions, $r = .50$, 95% CI [.19, .71].

Finally, we conducted a comparison of effect sizes across conditions an the estimation of their confidence intervals. Figure 4 offers a forest plot of these parameter and interval estimations.

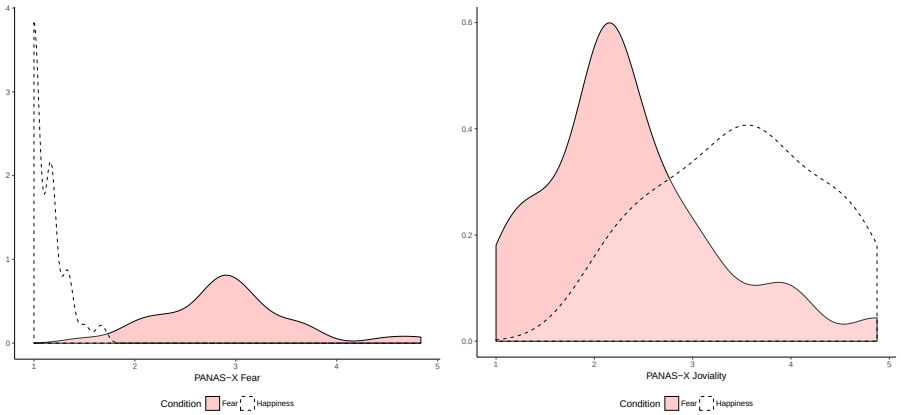
We observed that even if the manipulations yielded large and very large effects, the effect on password strength was trivial. The margin of error on the effect size estimation was less than half standard deviation.

Order Effects. Dependent samples t-tests of fear and joviality by the order of conditions showed no statistically significant difference, $ps > .45$ and $ps > .25$ respectively. We note that the differences of fear by order were neither normally nor symmetrically distributed, by which we used a dependent-samples Sign Test for the corresponding analysis.

Considering the impact of the order on the password strength is interesting in itself, because in the first password choice participants made an initial account registration, in the second password choice the participants made a password reset after an incident. The differences of zxcvbn log10 guesses fulfilled the assumptions (no outliers, normality) for a dependent-samples t-test.

zxcvbn log10 guesses were not statistically significantly different by condition order, $t(33) = -1.36, p = .183$, $g_{\mathrm{av}} = 0.28$, 95% CI [−0.13, 0.7]. The mean password strength for the first password (registration) was $M_{\mathrm{F,1st}} = 6.79$. The mean password strength for the second password (renewal) was $M_{\mathrm{F,2nd}} = 6.43$.

6.1.4 Correlation. We found that the dependent variable zxcvbn log10 guesses was not statistically significantly correlated with either fear or joviality. Table 6 offers an overview of the pair-wise correlations.



(a) PANAS-X fear **(b) PANAS-X joviality**

Figure 3: Density plots by manipulation check for Study 1: Fear.

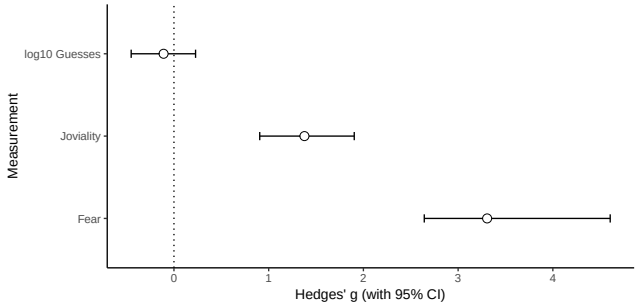


Figure 4: Comparison of the effects of the manipulations vis-'a-vis of the effect of fear on password choice. (Hedges' g_{av}, 95% Confidence Intervals).

Table 6: Correlations of fear vs. password strength in zxcvbn log10 **guesses [95% Confidence Intervals].**

	PX fear	PX joviality
PX fear		
PX joviality	-0.52*** [-0.67, -0.32]	
log10 Guesses	-0.03 [-0.27, 0.21]	0.08 [-0.16, 0.31]

6.2 Study 2: Effect of Stress

6.3 Data Preparation

We have analyzed the data for univariate outliers with the Outlier Labeling Rule as well as multivariate outliers with the Mahalanobis distance D^2. We found two cases with extreme values of zxcvbn log10 guesses greater than 16 and $D^2 > 19$. We decided to cap the outlying values with the 95th percentile, instead of removing the cases altogether.

6.3.1 Descriptives. Table 7 shows the means and standard deviations of the stress study.

6.3.2 Manipulation Check: SSSQ and STAI.

Table 7: Descriptive statistics of Study 2: Stress.

(a) Elicited Affect: Stress

	SQQQ		STAI	zxcvbn
	stress	distress	state_anxiety	log10 Guesses
M	71.50	16.78	38.50	7.96
SD	8.43	5.10	10.61	2.49

(b) Control

	SQQQ		STAI	zxcvbn
	stress	distress	state_anxiety	log10 Guesses
M	65.58	12.76	31.58	7.95
SD	9.15	4.04	7.07	2.35

Assumptions. For the Short State Stress Questionnaire (SSSQ) our analysis did not vouch for the capping of any outliers. The differences of the total stress values were normally distributed, Shapiro-Wilk $W = 0.98, p = .580$.

For SSSQ Distress, we observed two outliers, yet not extreme enough to vouch for capping or removal. The differences between distress scores across conditions were normally distributed, Shapiro-Wilk $W = 0.97, p = .158$.

For State-Trait Anxiety (STAI), there were no outliers. The difference between the state anxiety scores of both conditions were normally distributed, Shapiro-Wilk $W = 0.96, p = .081$.

Success of the Stress Manipulation. All three measurements, SSSQ stress and distress as well as STAI state_anxiety showed that the stress manipulations were successful. We offer an overview of the distributions of the treatments by measurements in Figure 5.

Participants in the stress condition showed statistically significantly more overall stress than in the control condition, $t(49) = 6.69, p_{MC(5)} < .001$, Hedges' $g_{av} = 0.66$, 95% CI [0.43, 0.91].

They exhibited statistically significantly more distress than in the control condition, $t(49) = 10.78, p_{MC(5)} < .001$, Hedges' $g_{av} = 0.87$, 95% CI [0.64, 1.11].

Furthermore, they exhibited statistically significantly more state anxiety than in the control condition, $t(49) = 6.46, p_{\text{MC}(5)} < .001$, Hedges' $g_{\text{av}} = 0.88$, 95% CI [0.59, 1.17].

As a consequence, we reject the null hypothesis $\text{H}_{\text{mc,S},0}$.

6.3.3 Password Strength.

Assumptions. While we detected one outlier, we decided not to cap it as it was close to the inner fence. The differences between zxcvbn log10 guesses between conditions were normally distributed, $W = 0.96, p = .094$.

Difference Between Conditions. There was no statistically significant mean difference log10 number of guesses between stress and control condition, $t(49) = 0.04, p = .971$, Hedges' $g_{\text{av}} = 0.01$, 95% CI $[-0.31, 0.33]$. Hence, we failed to reject the null hypothesis $\text{H}_{\text{S},0}$.

There was a statistically significant correlation between zxcvbn log10 guesses in the stress and control condition, $r = .33$, 95% CI [.06, .56].

We include a forest plot of the standardized mean difference of password strength under stress in Figure 6. Even though the manipulation caused stress with medium to large effect size, the effect size observed on password strength in log10 guesses is 0, where the confidence interval brackets it to an at most small effect.

Order Effects. Having checked the assumptions, we computed dependent-samples t-tests on stress, distress and state_anxiety by the order of experiment and control condition. There were no statistically significant order effects, $ps > .30$.

In first password choice participants made an initial account registration, in the second password choice the participants made a password reset after a security incident. zxcvbn log10 guesses were not statistically significantly different by condition order, $t(49) = -0.86, p = .397$, $g_{\text{av}} = 0.17$, 95% CI $[-0.22, 0.56]$.

The mean password strength for the first password (registration) was $M_{\text{S, 1st}} = 7.7511819$. The mean password strength for the second password (renewal) was $M_{\text{S, 2nd}} = 8.1579705$.

Difference by Reuse. We analyzed the difference in password strength depending on whether participants chose a new password or reused (a variant of) an old one. Under observation of the corresponding assumptions, computed independent-samples Welch t-tests. The zxcvbn log10 guesses were not statistically significantly different by reuse for either the experiment or the control condition respectively, EXP: $t(25.26) = 1.17, p = .253$, Hedges' $g = -0.26$, 95% CI $[-1.02, 0.5]$ and CTRL: $t(9.34) = -0.66, p = .528$, Hedges' $g = 0.27$, 95% CI $[-0.49, 1.02]$.

Consequently, we failed to reject the null hypothesis that reuse of passwords has no impact on the password strength.

6.3.4 Correlation. We found that there was no statistically significant correlation between the measurements of stress and the password strength in log10 guesses. Table 8 contains the overall correlation matrix.

6.4 Stress × Cognitive Load Interaction

We observed indications of a disordinal/cross-over interaction between the experiment condition and task load. Figure 7 offers an interaction diagram illustrating the situation.

We conducted a repeated-measures mixed-effects analysis with TLX Mental Demand and the Experiment Condition as fixed effects. We included a random intercept and distress slope with the subject as the context.

Compared to the intercept model, the model under consideration was on the borderline, but not statistically significant under a Likelihood-Ratio test, $\chi^2(6) = 12.589, p = .050 \not< .05$.

Using the multiple correlations R^2 between the original data and the fitted values as an estimate of overall effect size, we obtain $R^2 = 68.6\%$. Figure 8 illustrates the fit of the model.

In terms of model diagnostics, we found that the residuals were largely normally distributed, with slight deviations on the tails. We perceived the fitted-residual distribution as largely homoscedastic, even if with a linear bias hinting at a hidden variable.

We offer an overview of the model's coefficient estimates in Figure 9.

The impact of the Condition × TLX Mental Demand interaction was statistically significant, $F(1, 47) = 4.868, p = .032$, 0.25, 95% CI [0.03, 0.47].

We observe a statistically significant intercept estimate, $F(1, 49) = 899.161, p < .001$, 8.31, 95% CI [7.55, 9.07].

The main fixed effects were not statistically significant. Condition: $F(1, 47) = 0.004, p = .948$; TLX Mental Demand: $F(1, 47) = 0.261, p = .612$.

We observe the following cross-over interaction: In the control condition, that is when the participants completed non-stressing tasks, participants who reported low mental demand chose passwords with greater mean log10 guesses than participants reporting high mental demand.

However, in the experiment condition, when the participants completed stressful tasks, participants who reported high mental demand on average chose better passwords than participants who reported low mental demand.

We note that the companion analysis report [21] also contains an analysis of the three-way interaction Condition × TLM_Mental × Reuse. The mixed-effects model of the three-way interaction was not statistically significant, $\chi^2(10) = 17.941, p = .056$. Said analysis is congruent with stress and cognitive load having a stronger negative impact on password strength if the user chooses a new password.

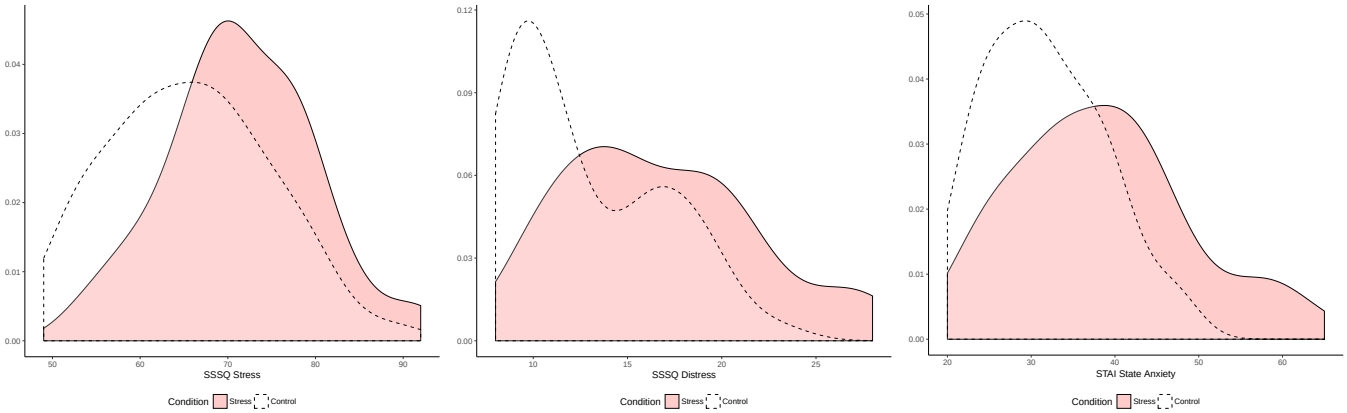
6.5 Meta Analysis of Order Effects

A meta analysis of the order effects across both studies showed an overall effect in Hedges' $g_{\text{av}} = 0.219$, 95% CI $[-0.061, 0.499]$. The summary effect was not statistically significant though, $p = .125$.

Cochran's Q-test was not statistically significant at $\alpha = .1$, $Q(1) = 0.156, p = .693$. We observed a heterogeneity of $I^2 = 0\%$.

6.6 Network Meta Analysis

We conducted a network meta analysis to put the effects of different studies in relation. We considered the 2016 LASER paper by Groß et al. [22], which considered the effect of cognitive effort and depletion on password choice.



(a) SSSQ stress **(b) SSSQ distress** **(c) STAI state_anxiety**

Figure 5: Density plots by manipulation check for Study 2: Stress.

Table 8: Correlations of stress vs. password strength in zxcvbn log10 guesses [95% Confidence Intervals].

	Total Stress	Distress	Anxiety
Total Stress			
Distress	0.64*** [0.51, 0.75]		
State Anxiety	0.35*** [0.16, 0.51]	0.64*** [0.51, 0.74]	
log10 Guesses	-0.02 [-0.21, 0.18]	0.08 [-0.12, 0.27]	0.05 [-0.14, 0.25]

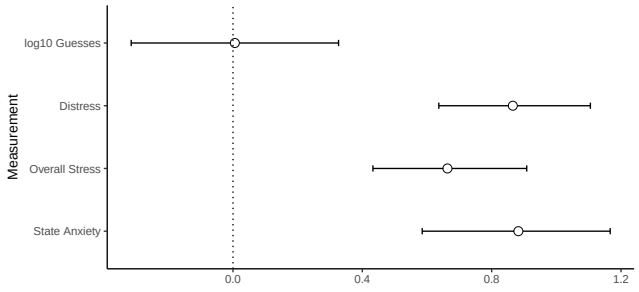


Figure 6: Comparison of the effects of the manipulations vis-'a-vis of the effect of stress on password choice. (Hedges' g_{av}, 95% Confidence Intervals).

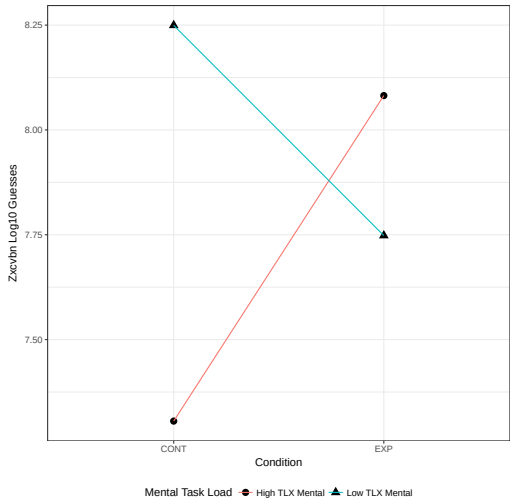


Figure 7: Interaction TLX Mental level by Condition.

For the network meta analysis, we coded Groß et al.'s categories "Undepleted," "Effortful" and "Depleted" conditions as three conditions of the same study, noting that, in fact, these categories are derived from the study's manipulation check on depletion level.

We coded the "Fear" and "Happiness" conditions of Study 1 of this paper, such that "Happiness" is mapped onto "Undepleted." Similarly, we coded the "Stress" and "Control" conditions of Study 2 of this paper, such that "Control" is mapped onto "Undepleted."

We display an overview of the resulting network meta analysis in Figure 11. The meta analysis is based on standard mean differences, Hedges' g in case of Gross et al. and Hedges' g_{av} in case of this paper. Figure 11a yields a forest plot of the results, while Figure 11b shows the network of treatment relations.

We observe that the effects associated with the "Effortful" and "Depleted" categories of Groß et al. [22] are maintained at large and medium effect sizes. The treatments fear and stress only yield trivial effect sizes. Consequently, stress is not supported as an alternative explanation for reported effects of cognitive effort and depletion.

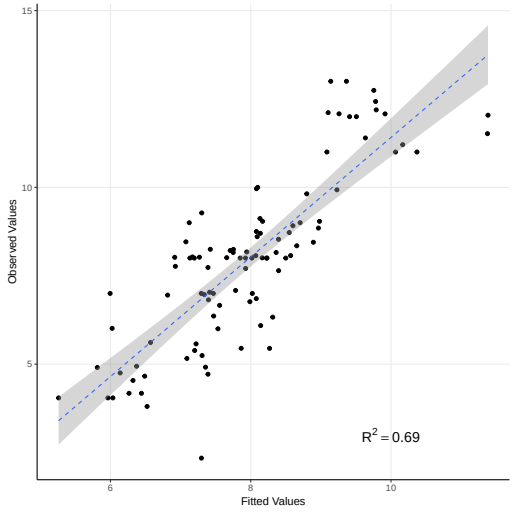


Figure 8: Fit of the stress–mental demand interaction model.

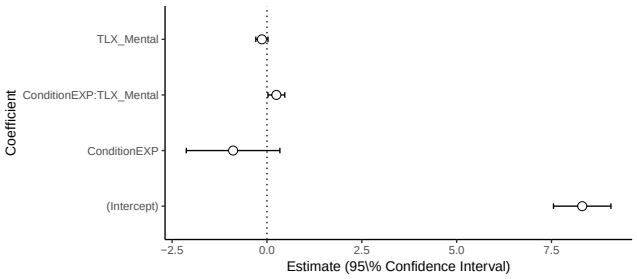


Figure 9: Interaction Model Coefficient Estimates.

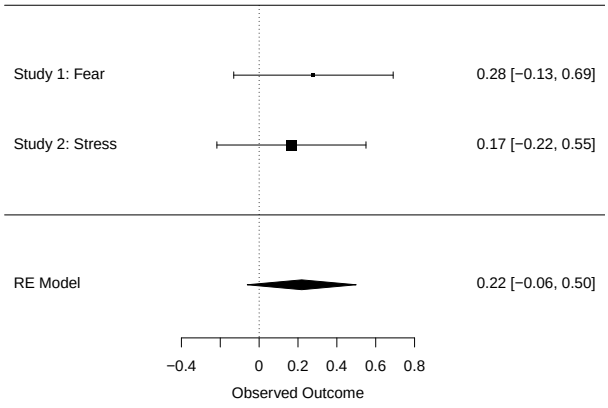


Figure 10: Forest Plot of order effects

7 DISCUSSION

7.1 Incidental fear and stress were successfully induced.

We induced incidental fear and stress, that is, an affective state not related to the password choice scenario. Consistently yielding large to very large effect sizes, the different induction techniques (affect stimulus videos, stress battery) were shown to have worked well.

Clearly, task-related fear and stress should more likely to impact password strength, as pursued in research on fear appeals. For instance, participants could be exposed to a news article describing the negative impact of identity theft or a password breach. Such an experiment setup would vouch for an analysis with the Protection Motivation Theory (PMT) [17, 42, 56].

7.2 Incidental fear and stress is likely to have at most a small negative effect on password choice.

While the 95% confidence intervals on the effect of fear and stress on password strength bracketed the effects as small, the effect of incidental stress corrected for other predictors was negative, similarly to the effect of incidental fear. Hence, future research may aim at pinpointing a negative influence of incidental stressors.

At the same time, we found a statistically significant interaction between stress and mental demand, which asks for further investigation seeking to isolate both conditions.

7.3 Whether reusing an existing password or creating a new one serves the user better may be situational.

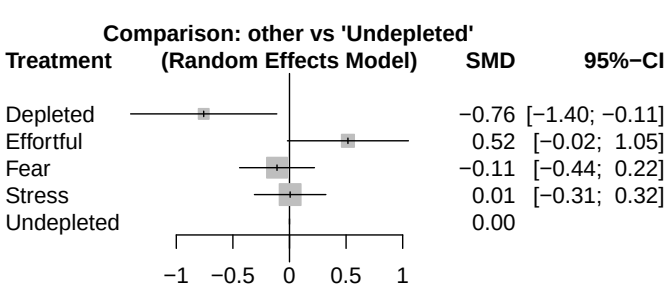
While a three-way interaction model including password reuse [21] was not statistically significant, we observe weak evidence that newly created passwords would be weaker when the user is either stressed or under mental demand, and stronger under baseline conditions.

While recommendations that users should create new passwords when they are rested as well as be allowed to rely on variants of prior passwords when they are stressed or depleted seem plausible, this area requires further investigation for a conclusive result.

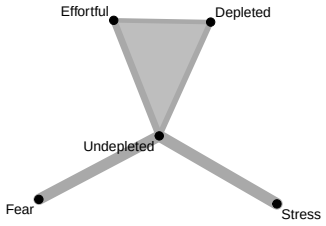
7.4 Cognitive effort and depletion are maintained as strong effects on password choice.

Groß et al. hypothesized that stress could be an alternative explanation for the observed effects (a) that users under cognitive effort but not depleted created better passwords than the control group, and (b) that users reporting high depletion create worse passwords than the control group.

Having induced incidental fear and stress and compared the results in a network meta analysis, we have not found evidence that these factors cause an effect of similar magnitude as the the cognitive effort and depletion, reported in Gross et al. [22]. Hence, cognitive effort and depletion are still plausible explanations of the observed differences in password strength.



(a) Forest Plot of Treatment Effects



(b) Network Graph

Figure 11: Network Meta Analysis of LASER'2016 [22] as well as this paper's Study 1: Fear and Study 2: Stress.

7.5 When users are asked to renew their password because of a security incident, their password strength may improve.

Having considered the meta analysis of differences between password choice in a first-time registration and in a renewal after a security incident, we found evidence of a consistent effect between studies that the password strength in the renewal condition is slightly greater than in the first-time registration.

In terms of magnitude, we find that the greater log10 number of guesses during renewal after a reported security incident makes for a small effect size, $g_{\text{av}} = 0.219$, 95% CI $[-0.061, 0.499]$.

This yields an early indication that task-related fear (induced by a security incident message) is causing a change in user behavior towards a protection motivation. Hence, we expect that a future experiment constructed to test the full Protection Motivation Theory (PMT) in a password choice scenario will yield conclusive result.

Given the small effect size estimates, however, neither the individual studies nor the meta analysis had enough power to reject the null hypothesis with statistical significance. Hence, this asks for further investigation.

7.6 Limitations

7.6.1 Generalizability. The participants were largely recruited from university students, limiting generalizability.

In terms of ecological validity, the studies created a scenario in which actual private information of participants (personality traits, stress and anxiety data) was stored on a Web site. Having been made aware of the sensitivity of such personal data, the participants' incentive to protect the data was similar to real life.

The participants were exposed to a diversion in that they were only informed after the experiment that the research aim included the password strength and in that they were misled under the pretext of a security incident to change their password. Consequently, the first and the second run of the password choice trial were different by design in terms of "first registration" vs. "password change after incident".

7.6.2 Constraints on Bounding Small Effects. We note that an within-subjects experiment to establish a lower bound on the impact of fear or stress in password strength would need a considerable number of participants. For 95% *a priori* power on a dependent-samples t-test for standardized mean difference of Cohen's $d = 0.1$, one would need a sample size of 1300.

8 CONCLUSION

This is the first work to investigate the effects of incidental fear and stress on password choice. It is the first to estimate the magnitude of such effects across studies on the influence of the user's current cognitive and affective state on password decision making.

As future work, the two studies yield an observation on the effect of fear appeals in password choice. There were first indications that the message of a security incident caused participants to choose a stronger password with a small effect size. This vouches for further investigation, for instance using the full Protection Motivation Theory (PMT) [17, 42, 56] as a foundation.

ACKNOWLEDGEMENTS

We are grateful for the discussions with Kovila Coopamootoo, especially on earlier work on cognitive effort, on incidental and integral affect, as well as on the Protection Motivation Theory. We are grateful for the discussions with Uchechi Phyllis Nwadike, especially on the effects of incidental fear, sadness and happiness on privacy decision making [40]. We appreciated discussions with Roy Maxion on experimentation on stress. This work was in parts supported by the ERC Starting Grant CASCAde (GA n°716980).

REFERENCES

[1] Anne Adams and Martina Angela Sasse. 1999. Users are not the enemy. *Commun. ACM* 42, 12 (1999), 40–46.
[2] James Algina and HJ Keselman. 2003. Approximate confidence intervals for effect sizes. *Educational and Psychological Measurement* 63, 4 (2003), 537–553.
[3] R.F. Baumeister, E. Bratslavsky, E. Muraven, and D.M. Tice. 1998. Ego depletion: is the active self a limited resource? *Personality and social psychology* 74 (1998), 1252–1265.
[4] Roy F Baumeister, Kathleen D Vohs, and Dianne M Tice. 2007. The strength model of self-control. *Current directions in psychological science* 16, 6 (2007), 351–355.
[5] Joseph Bonneau. 2012. The science of guessing: analyzing an anonymized corpus of 70 million passwords. In *Security and Privacy (SP), 2012 IEEE Symposium on*. IEEE, 538–552.
[6] Joseph Bonneau, Cormac Herley, Paul C Van Oorschot, and Frank Stajano. 2012. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *Security and Privacy (SP), 2012 IEEE Symposium on*. IEEE, 553–567.
[7] Joseph Bonneau and Sören Preibusch. 2010. The Password Thicket: Technical and Market Failures in Human Authentication on the Web.. In *WEIS*.

[8] Scott R Boss, Dennis F Galletta, Paul Benjamin Lowry, Gregory D Moody, and Peter Polak. 2015. What do users have to fear? Using fear appeals to engender threats and fear that motivate protective security behaviors. (2015).

[9] William E. Burr, Donna F. Dodson, and W. Timothy Polk. 2004. *Electronic Authentication Guideline.* NIST Special Publication 800-63. NIST.

[10] James A Coan and John JB Allen. 2007. *Handbook of emotion elicitation and assessment.* Oxford university press.

[11] Harris Cooper, Larry V Hedges, and Jeffrey C Valentine. 2009. *The handbook of research synthesis and meta-analysis.* Russell Sage Foundation.

[12] Geoff Cumming. 2013. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* Routledge.

[13] Anupam Das, Joseph Bonneau, Matthew Caesar, Nikita Borisov, and XiaoFeng Wang. 2014. The Tangled Web of Password Reuse.. In *NDSS*, Vol. 14. 23–26.

[14] Richard J Davidson, Klaus R Sherer, and H Hill Goldsmith. 2009. *Handbook of affective sciences.* Oxford University Press.

[15] Dinei Florencio and Cormac Herley. 2007. A large-scale study of web password habits. In *Proceedings of the 16th international conference on World Wide Web.* ACM, 657–666.

[16] Dinei Florêncio, Cormac Herley, and Paul C Van Oorschot. 2014. Password Portfolios and the Finite-Effort User: Sustainably Managing Large Numbers of Accounts.. In *USENIX Security Symposium.* 575–590.

[17] Donna L Floyd, Steven Prentice-Dunn, and Ronald W Rogers. 2000. A meta-analysis of research on protection motivation theory. *Journal of applied social psychology* 30, 2 (2000), 407–429.

[18] Tom Fordyce, Sam Green, and Thomas Groß. 2017. *Investigation of the Effect of Fear and Stress on Password Choice (Extended Version).* Technical Report TR-1517. Newcastle University.

[19] Martin J Gardner and Douglas G Altman. 1986. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)* 292, 6522 (1986), 746–750.

[20] James J Gross and Robert W Levenson. 1995. Emotion elicitation using films. *Cognition & emotion* 9, 1 (1995), 87–108.

[21] Thomas Groß. 2017. *Analysis Report – Investigation of the Effect of Fear and Stress on Password Choice.* OSF Report https://osf.io/3cd9h/. Open Science Framework.

[22] Thomas Groß, Kovila Coopamootoo, and Amina Al-Jabri. 2016. Effect of Cognitive Depletion on Password Choice. In *Learning from Authoritative Security Experiment Results (LASER'16)*, Sean Peisert (Ed.).

[23] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage Publications, 904–908. Issue 9.

[24] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology* 52 (1988), 139–183.

[25] William S Helton. 2004. Validation of a short stress state questionnaire. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 48. SAGE Publications Sage CA: Los Angeles, CA, 1238–1242.

[26] William S Helton and Katharina Näswall. 2015. Short Stress State Questionnaire: Factor structure and state change assessment. *European Journal of Psychological Assessment* 31, 1 (2015), 20.

[27] Peter Hoonakker, Nis Bornoe, and Pascale Carayon. 2009. Password authentication from a human factors perspective. In *Proc. Human Factors and Ergonomics Society Annual Meeting*, Vol. 53. SAGE Publications, 459–463.

[28] Daniel Kahneman. 1973. *Attention and effort.* Citeseer.

[29] Patrick Gage Kelley, Saranga Komanduri, Michelle L Mazurek, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Julio Lopez. 2012. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *Security and Privacy (SP), 2012 IEEE Symposium on.* IEEE, 523–537.

[30] Megan M Kelly, Audrey R Tyrka, George M Anderson, Lawrence H Price, and Linda L Carpenter. 2008. Sex differences in emotional and physiological responses to the Trier Social Stress Test. *Journal of behavior therapy and experimental psychiatry* 39, 1 (2008), 87–98.

[31] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H Hellhammer. 1993. The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology* 28, 1-2 (1993), 76–81.

[32] Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology* 4 (2013).

[33] Michael Lewis, Jeannette M Haviland-Jones, and Lisa Feldman Barrett. 2010. *Handbook of emotions.* Guilford Press.

[34] RN Li-Mei Liao and Mary G Carey. 2015. Laboratory-induced mental stress, cardiovascular response, and psychological characteristics. *Rev Cardiovasc Med* 16, 1 (2015), 28–35.

[35] Gerald Matthews, Lucy Joyner, Kirby Gilliland, SE Campbell, Shona Falconer, and Jane Huggins. 1999. Validation of a comprehensive stress state questionnaire: Towards a state big three. *Personality psychology in Europe* 7 (1999), 335–350.

[36] Karen A Matthews and Catherine M Stoney. 1988. Influences of sex and age on cardiovascular responses during stress. *Psychosomatic Medicine* 50, 1 (1988), 46–56.

[37] John D Mayer and Yvonne N Gaschke. 1988. The experience and meta-experience of mood. *Journal of personality and social psychology* 55, 1 (1988), 102.

[38] Binod Neupane, Danielle Richer, Ashley Joel Bonner, Taddele Kibret, and Joseph Beyene. 2014. Network Meta-Analysis Using R: A Review of Currently Available Automated Packages. *PLOS ONE* 9, 12 (12 2014), 1–17. https://doi.org/10.1371/journal.pone.0115065

[39] Raymond S Nickerson. 2000. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods* 5, 2 (2000), 241.

[40] Uchechi Nwadike, Thomas Groß, and Kovila PL Coopamootoo. 2016. Evaluating Users' Affect States: Towards a Study on Privacy Concerns. In *Privacy and Identity Management. Facing up to Next Steps.* Springer, 248–262.

[41] Ellen Peters, Daniel Västfjäll, Tommy Gärling, and Paul Slovic. 2006. Affect and decision making: A "hot" topic. *Journal of Behavioral Decision Making* 19, 2 (2006), 79–85.

[42] Ronald W Rogers. 1983. Cognitive and psychological processes in fear appeals and attitude change: A revised theory of protection motivation. *Social psychophysiology: A sourcebook* (1983), 153–176.

[43] Jonathan Rottenberg, Rebecca D Ray, and James J Gross. 2007. Emotion elicitation using films. *Handbook of emotion elicitation and assessment* (2007), 9–28.

[44] Gerta Rücker. 2012. Network meta-analysis, electrical networks and graph theory. *Research Synthesis Methods* 3, 4 (2012), 312–324.

[45] Robert AC Ruiter, Loes TE Kessels, Gjalt-Jorn Y Peters, and Gerjo Kok. 2014. Sixty years of fear appeal research: Current state of the evidence. *International journal of psychology* 49, 2 (2014), 63–70.

[46] James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110, 1 (2003), 145.

[47] Martina Angela Sasse, Sacha Brostoff, and Dirk Weirich. 2001. Transforming the 'weakest link'—a human/computer interaction approach to usable and effective security. *BT technology journal* 19, 3 (2001), 122–131.

[48] Guido Schwarzer, James R. Carpenter, and Gerta Rücker. 2015. *Network Meta-Analysis.* Springer International Publishing, Cham, 187–216. https://doi.org/10.1007/978-3-319-21416-0_8

[49] Hans Selye. 1974. Stress without distress. *New york* (1974), 26–39.

[50] Charles Donald Spielberger, Richard L Gorsuch, and Robert E Lushene. 1970. Manual for the state-trait anxiety inventory. (1970).

[51] Karl Halvor Teigen. 1994. Yerkes-Dodson: A law for all seasons. *Theory & Psychology* 4, 4 (1994), 525–547.

[52] Dianne M Tice, Roy F Baumeister, Dikla Shmueli, and Mark Muraven. 2007. Restoring the self: Positive affect helps improve self-regulation following ego depletion. *Journal of Experimental Social Psychology* 43, 3 (2007), 379–384.

[53] David Watson and Lee Anna Clark. 1999. *The PANAS-X: Manual for the positive and negative affect schedule – expanded form.* Technical Report. University of Iowa, Department of Psychology.

[54] Rainer Westermann, GUNTER Stahl, and F Hesse. 1996. Relative effectiveness and validity of mood induction procedures: analysis. *European Journal of social psychology* 26 (1996), 557–580.

[55] Dan Lowe Wheeler. 2016. zxcvbn: Low-budget password strength estimation. In *Proc. USENIX Security.*

[56] Kim Witte. 1992. Putting the fear back into fear appeals: The extended parallel process model. *Communications Monographs* 59, 4 (1992), 329–349.

[57] Moshe Zviran and William J. Haga. 1993. A comparison of password techniques for multilevel authentication mechanisms. *Comput. J.* 36, 3 (1993), 227–237.