An Empirical Investigation of Security Fatigue
The Case of Password Choice after Solving a CAPTCHA

Kovila P.L. Coopamootoo
Newcastle University
kovila.coopamootoo@ncl.ac.uk

Thomas Groß
Newcastle University
thomas.gross@ncl.ac.uk

M. Faizal R. Pratama
University of Derby

Abstract

Background. User fatigue or overwhelm in current security tasks has been called *security fatigue* by the research community [11, 24]. However, security fatigue can also impact subsequent tasks. For example, while the CAPTCHA is a widespread security measure that aims to separate humans from bots [26], it is also known to be difficult for humans [2]. Yet, to-date it is not known how solving a CAPTCHA influences other subsequent tasks.
Aim. We investigate users' password choice after a CAPTCHA challenge.
Method. We conduct a between-subject lab experiment. Three groups of 66 participants were each asked to generate a password. Two groups were given a CAPTCHA to solve prior to password choice, the third group was not. Password strength was measured and compared across groups.
Results. We found a significant difference in password strength across conditions, with $p = .002$, corresponding to a large effect size of $f = .42$. We found that solving a text- or picture-CAPTCHA results in significantly poorer password choice than not solving a CAPTCHA.
Conclusions. We contribute a first known empirical study investigating the impact of a CAPTCHA on password choice and of designing security tasks in a sequence. It raises questions on the usability, security fatigue and overall system security achieved when password choice follows another effortful task or is paired with a security task.

1 Introduction

The CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart [26]) is an important web security measure that differentiates humans from bots. It is widely used across the web [20], including at websites with a high traffic flow such as Facebook, Twitter, LinkedIn, Reddit and various email providers. Balancing both security and usability of CAPTCHAs continues to be a challenge [30, 7]. CAPTCHAs are known to be difficult for humans [2] and to pose usability issues [29].

Security research shows that on the one hand, users are fatigued and frustrated with security [11]. Literature quotes a *compliance budget* where employees comply either when there is no extra effort or after weighing the cost and benefits of extra effort [1]. In addition, a threshold exists beyond which it gets too hard and burdensome for users to maintain security [11]. Fatigue impacts current tasks with users being desensitized to security or rejecting security [24].

On the other hand, there is a suspicion that priming moderate effort can enhance security decisions such as for password choice. For example, Groß et al. [12] found that while password strength is weakest when users are cognitively depleted, moderate effort exertion is beneficial for password strength.

When registering an account, users are often asked to solve a CAPTCHA and to choose a password. Although online account registration forms often present the CAPTCHA challenge after password choice, unclear guidance into the data entry sequence or form refresh when an incorrect CAPTCHA is entered lead to a situation where the password is chosen after solving the CAPTCHA. In addition, when the Tor Anonymizer is detected, users are systematically asked to solve a CAPTCHA before they can register. Therefore the question arises whether the effort required to solve a CAPTCHA impacts password choice. While a handful of research have studied a link between cognitive effort and password choice or password management [12, 17, 10], for empirical investigation of the impact of effort previously spent on password choice, we are only aware of Groß et al. [12].

Research Question. We investigate the main RQ *"How does solving a CAPTCHA before creating a pass-*

word influence password choice?" We reproduce experimental design components of Groß et al. [12] and use validated methods from cognitive psychology to measure effort [25], stress[22, 23] and cognitive load [13].

Contribution. This paper contributes the first empirical investigation of the impact of security fatigue (experienced in a first task) on a subsequent security task. Our findings indicate that engaging in a CAPTCHA influences password choice. Solving a text- or picture-CAPTCHA lead to weaker password than not solving a CAPTCHA with a large effect size of 0.42.

Outline. In the rest of the paper, we first provide background research, followed with the study aims. We describe a pre-study. We then follow the main study methodology and provide the results before the discussion, limitations and conclusion.

2 Background

2.1 Security Fatigue

Previous research suggests that users often perceive security as a barrier that interferes with their productivity [11]. Subsequently, the term *security fatigue* was coined to describe the threshold of acceptance beyond which it gets too burdensome for users to maintain security [11]. In this sense, security fatigue describes an interference to current tasks and an additional step to be taken. *Security fatigue* has also been used to describe users' weariness or reluctance to experience anymore of something [24]. In particular, they reported that participants are tired, turned off and overwhelmed by security. The term has been used to describe user behavior both in the workplace [11] and for the general public [24].

2.2 CAPTCHA

A CAPTCHA is a program that can generate and grade tests that most humans can pass, but that current computer programs cannot [26]. By differentiating humans from bots, they are used for security applications such as preventing bots from continuous auto-voting in online polls, -registering to email accounts or preventing dictionary attacks [26]. Text-based CAPTCHAs require users to type alphanumeric characters from a distorted image, where popular ones include reCAPTCHA [27, 21] and BaffleText [5] whereas image-recognition CAPTCHAs are based on image problems, where examples include ASIRRA [8] and reCAPTCHA [21].

2.2.1 Usability & Security

While research propose that good CAPTCHAs ought to be both usable by humans and strong in resisting adversarial attacks [29, 7], CAPTCHAs are often difficult for humans [2, 4, 29, 9]. User reports on the perception, preferences and use of CAPTCHA [9] show that only every other user solves a CAPTCHA at first try, with character distortion named as the main obstacle.

Early CAPTCHAs have been broken by object recognition algorithms [18, 4] and segmentation [30, 7]. Yan and El Ahmad [30] exploited flaws in a word-image scheme via simple attacks and found that it was easy to separate foreground text from the background and that the scheme was vulnerable to segmentation attacks and dictionary attacks when English words was used. Bursztein et al. [3] provided an enhancement to the process of attacking text CAPTCHAs and they proposed randomizing the CAPTCHA length and individual character size, creating a wave shape and collapsing or overlaid lines, for improved protection against attacks. There are also claims of breaking the latest of Google's image reCAPTCHA [21].

2.3 Text Password

Text passwords are the cheapest and most commonly used method of computer authentication. However, a large proportion of users are frustrated when forced to comply to password policies such as monthly reset [15]. Effort and tiredness to a state of cognitive depletion causes users to choose weaker passwords [12], providing an indication that effort is necessary for the creation of strong passwords.

3 Aim

We investigate the main RQ *"How does solving a CAPTCHA before creating a password influence password choice?"*.

3.1 Impact on Password Strength

Password strength varies according to cognitive state, that is whether the user is depleted or fresh [12].

Research Question 1 (RQ⊨P). *How does the strength of a password chosen after solving a CAPTCHA differ from not solving a CAPTCHA?*
$H_{P,0}$: *Solving a CAPTCHA does not impact password strength*
$H_{P,1.1/P,1.2}$: *Solving a [text-CAPTCHA/picture-CAPTCHA] causes weaker passwords than in the CONT condition.*

3.2 Effort Exerted

In Section 2.2, we reviewed literature exposing the difficulties of solving CAPTCHAs. These difficulties entail user effort.

Research Question 2 (RQ⊨E)**.** *How does the overall effort of solving a CAPTCHA and choosing a password differ from only choosing a password?*
$\mathsf{H}_{\mathsf{E},1.1/\mathsf{E},1.2}$*: Solving a [text-CAPTCHA/picture-CAPTCHA] causes exertion of more effort than not solving any CAPTCHA.*
$\mathsf{H}_{\mathsf{E},1.3}$*: Solving a text-CAPTCHA causes exertion of more effort than solving a picture-CAPTCHA.*

3.3 Performance

The text and picture-CAPTCHAs can pose difficulties that affect performance differently, for example while the distortions of text-CAPTCHAs are problematic for the user, recognition or picture quality might do so for picture-CAPTCHA.

Research Question 3 (RQ⊨F)**.** *How does the type of CAPTCHA impact the time spent, success rate and results checking rate?*
$\mathsf{H}_{\mathsf{T},1}$*: Solving a text-CAPTCHA requires more time.*
$\mathsf{H}_{\mathsf{S},1}$*: Solving a text-CAPTCHA has lower success rate.*
$\mathsf{H}_{\mathsf{R},1}$*: Solving a text-CAPTCHA has a higher results checking rate.*

4 Pre-Study

Before the main study reported in this paper, we designed a pre-study to compare password strength across two effort inducing conditions.

4.1 Aim

Research Question 4. *How does password choice differ between different conditions of effort, in particular across control, text-CAPTCHA and 2-digit multiplication?*

4.2 Method

4.2.1 Participants

Participants were recruited on the university campus via adverts and flyers. They were paid a time compensation of \$6.5. Sample size was $N = 40$, with 17 women, mean age 26.75 years ($SD = 7.540$). 9 participants were from a computer science background.

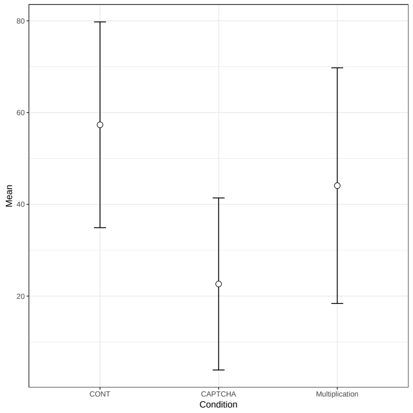

Figure 1: Confidence Intervals of the means of password strength score by condition. (Pre-Study)

4.2.2 Procedure

The procedure consisted of (a) a pre-task questionnaire for demographics, (b) a mood questionnaire, (c) a puzzle manipulation, (d) a password entry for a mock-up GMail registration, (e) a mood questionnaire and (f) a debriefing questionnaire.

We choose two puzzles: an example of the widely used text-CAPTCHA and a 2-digit multiplication. We choose a 2-digit multiplication because it is known in cognitive psychology to consume effort [16]. Solving mental multiplication problems has been shown to engage cognitive effort in particular 2-digit multiplication [14]. In this task we asked participants to solve 48 x 97 (ensuring that one of the numbers was a prime number, since prime numbers ensure shortcuts are not being used). We provide more details of the text-CAPTCHA in Section 5.2.2, which is also used for the main study.

We designed a between-subject study where participants were randomly assigned to one of the three groups. We ended up with 12 participants in the CONT, 14 in the text-CAPTCHA condition and 14 in the 2-digit multiplication condition.

4.3 Results and Discussion

Similar to Groß et al. [12], we measure password strength via password meter with NIST amendments. We computed a one-way ANOVA with the password strength score as dependent variable. There was no statistically significant effect of the experiment condition on the password strength score, $p = .074 > .05$.

Considering the intervals plot of Figure 1, we observe that the confidence intervals of the conditions CONT and CAPTCHA barely overlap, asking for further investigation. In terms of effect size, we see an effect size of Hedges' $g = 0.99$ $[0.17, 1.81]$.

5 Main Study Method

5.1 Participants

Participants were recruited on campus via adverts and flyers, and the experiment was run in a dedicated quiet zone. Participants were paid a time compensation of $6.5 for completing the experiment, which lasted between 10-15 minutes. The sample consisted of university students, $N = 66$, of which 30 were women. The mean age was 21.79 years ($SD = 3.223$). Participants were not from a computer science background, 38 were local nationals, and 63 reported undergraduate education.

5.2 Procedure

We designed a between-subject experiment, via experimental design guidelines [6]. We induce the independent variable (IV) effort, with three levels: CONT, text-CAPTCHA and picture-CAPTCHA, further described in Section 5.2.2.

The procedure consisted of (a) a pre-task questionnaire for demographics, (b) a combined short stress and mood questionnaire, (c) a CAPTCHA manipulation, (d) a password entry for a mock-up GMail registration (a reproduction of [12]), (e) a combined full stress and mood questionnaire, (f) a questionnaire for task load and (g) a debriefing questionnaire. Figure 2 depicts the experiment design.

5.2.1 Block Randomization

We ensured that each condition had equal number of participants and that participants are randomly assigned across groups. We automated a random block assignment method and we ended up with an equal number of 22 participants in each condition.

5.2.2 Manipulation Tasks

Following the review in Section 2.2, we selected the character/text-CAPTCHA and image recognition/picture-CAPTCHA as the experimental conditions. We chose these two schemes because the text-CAPTCHA has been most popular and is still used by high traffic sites such as Facebook[1] while the picture-CAPTCHA is an option for the reCAPTCHA, which is the most used CAPTCHA according to online surveys [20]. In terms of security, both schemes are also known to suffer from security flaws and have been been broken by segmentation and machine learning attacks as seen in Section 2.2.

Text-CAPTCHA We generated a CAPTCHA image using Securimage PHP CAPTCHA[2], as shown in Figure 3. Securimage distorts the code and draws random lines over the image. We used a level of perturbation of 1.75 to induce effort. 1.75 is readable yet require some effort. The number of lines on the image was set to the default 5. This CAPTCHA was also used in the pre-study provided in the Appendix.

Picture-CAPTCHA The image reCAPTCHA challenge provides a sample image and 9 candidate images. It asks the user to select images similar to the sample [21], where the correct number of images vary between 2 to 4. In our picture-CAPTCHA condition, we tweaked the image reCAPTCHA process and asked participants to count the number of times a particular image appear, here the number of cats as shown in Figure 4. We estimated that the effort spent in clicking on all occurrences would be similar as counting the number of occurrences. Participants still have to recognize particular images, yet we maintain a similar user input (text entry) as in the text-CAPTCHA condition.

5.3 Measures

5.3.1 Password Strength

Similar to [12], we use password meter Web site[3] with NIST adjustments. In addition, we evaluated the zxcvbn password strength estimator [28]. Zxcvbn provides the number of guesses, $\log_{10}$ guesses and a zxcvbn score from 0 to 4.

5.3.2 Password Strategy and Re-Use

At the debrief, we asked participants if they re-used one of their existing passwords to register to the GMail account. We also queried for password strategy employed. We report these in the Appendix.

5.3.3 Brief Mood Inventory

As [12], we use a short form of a brief mood inventory (BMI). Because we merged the stress and BMI questionnaire, instead of a 5-point Likert-type items between 1 Disagree strongly and 5 Agree strongly, we used a the 4-point Likert of the stress questionnaire with items 1 Not at all, 2 Somewhat, 3 Moderately and 4 Very much.

5.3.4 Stress and Workload

The Spielberger State Trait Anxiety Inventory is one of the most used measures of anxiety and stress in psychol-

[1] https://www.facebook.com

[2] https://www.phpcaptcha.org

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

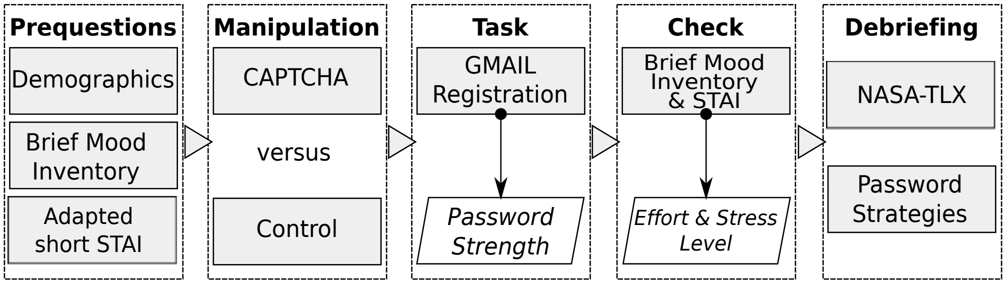

Figure 2: Overview of the experiment procedure. The experimental groups solved a CAPTCHA. The control group did not.

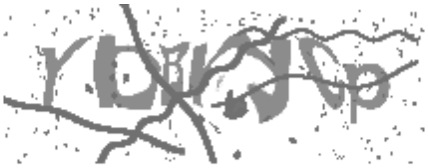

Figure 3: The text-CAPTCHA

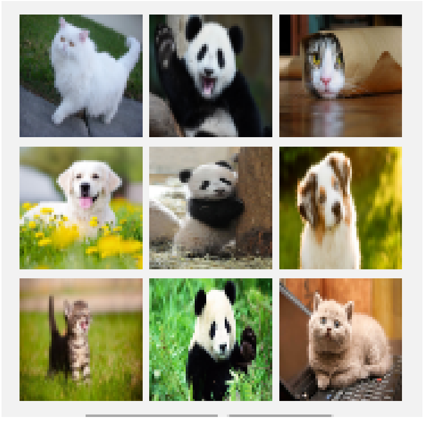

Figure 4: The picture-CAPTCHA

ogy [22, 23]. We chose the Y-1 questionnaire as measure of stress, with items towards how the participant felt in the experiment. In the post-task questionnaire, we included the full STAI.

NASA Task Load Index (NASA TLX) assesses mental workload via the dimensions of mental demand, physical demand, temporal demand, performance, effort and frustration [13].

5.3.5 Performance

Previous research recorded the time required [19] as well as the number of attempts required by participants to solve CAPTCHAs [9]. We also recorded the time taken in seconds to solve the CAPTCHAs, the final result entered by the participant and the number of times participants checked their results.

6 Results

All inferential statistics are computed at a significance level α of 5%. We estimate population parameters, such as standardized effect sizes of differences between conditions with 95% confidence intervals. A *confidence interval* is an interval estimate of a population parameter. The confidence level determines the frequency such confidence intervals would contain the population parameter if an infinite number of independent experiments were conducted.

6.1 Manipulation Check

6.1.1 Effort Exerted

We evaluate the null hypothesis $\mathsf{H}_{\mathsf{E},0}$: *Solving a CAPTCHA does not impact the effort exerted.* We calculate Diff_Tiredness (from the BMI in Section 5.3.3) as the difference in self-reported tiredness before the start of the CAPTCHA and after the registration.

A Kruskal-Wallis test showed that there was a statistically significant difference in Diff_Tiredness between the different conditions $\chi^2(2) = 12.736$, $p = .002 < .05$, with a mean rank Diff_Tiredness 26.64 for CONT, 44.00 for text-CAPTCHA and 29.86 for picture-CAPTCHA. We reject the null hypothesis $\mathsf{H}_{\mathsf{E},0}$.

We run 3 Mann-Whitney tests, with a Bonferroni-corrected significance level of $\alpha_{\mathsf{B}} = .0167$: (a) Diff_Tiredness was statistically significantly greater in the text-CAPTCHA condition than in the control condition, $U = 112.50$, $Z = -3.432$, $p = .001 < .0167$, $r = -.42$. This constitute a large effect size; (b) There was no significant difference in Diff_Tiredness between the picture-CAPTCHA condition and the control condition, $p = .556 > .0167$; (c) Diff_Tiredness was statistically significantly greater in the text-CAPTCHA condition than in the picture-CAPTCHA condition, $U = 140.50$, $Z = -2.577$, $p = .0100 < .0167$, $r = -.32$. We observe a medium to large effect size.

From these results, we conclude that the manipulation was successful in leading participants to exert more effort in the text-CAPTCHA condition than in the picture-CAPTCHA and CONT conditions.

6.1.2 Performance

We evaluate the null hypotheses $\mathsf{H}_{\mathsf{T},0}$/ $\mathsf{H}_{\mathsf{S},0}$/ $\mathsf{H}_{\mathsf{R},0}$: *A type of CAPTCHA does not impact the [time spent/success rate/results checking rate].* These DVs indicate participants' engagement and enable further evaluation of the success of the manipulation.

Time to solve CAPTCHA. We note that Levene's test for equality of variances across conditions is not significant with $p = .640 > .05$. With a two-tailed independent samples $t - test$, we find that participants in the text-based CAPTCHA condition ($M = 128.94$, $SD = 22.18$) have taken statistically significantly more completion time than participants in the picture-CAPTCHA condition ($M = 32.57$, $SD = 19.60$), $t(42) = 15.271$, $p < .001$. This gives an effect size of Hedges' $g = 4.52$ $[3.38, 5.64]$, a very large effect. We reject the null hypothesis $\mathsf{H}_{\mathsf{T},0}$.

Success at completing CAPTCHA. Although participants seem to have tried longer for the text-CAPTCHA, only five of them obtained a correct result; 20 did so for the picture-CAPTCHA. We run a χ^2 test, where we find a significant difference in correct results across the CAPTCHA conditions with χ^2(1, N = 44) = 23.21, $p < .001$. The odds of having a correct result in the picture-CAPTCHA were 68 times higher than in the text-CAPTCHA. We reject the null hypothesis $\mathsf{H}_{\mathsf{S},0}$.

Checking Results. We counted the number of times participants checked their results. Since the number of checks failed Levene's test for equality of variance across the CAPTCHA conditions, with $p = .002 < .05$, we opt for the non-parametric Mann-Whitney test. The number of checks was significantly larger in the text-CAPTCHA condition than in the picture condition, $U = 19.5$, $Z = -5.365$, $p = .000 < .005$, $r = -.66$. This refers to a large effect size. We reject the null hypothesis $\mathsf{H}_{\mathsf{R},0}$.

6.2 Stress and Workload

We investigate the overall STAI score and the difference between the five pre/post STAI items and find no significant difference across the experimental conditions. We investigate NASA-TLX's across the dimensions of mental demand, physical demand, temporal Demand, performance, effort, frustration and the overall TLX_Score. We find no significant difference across conditions. We believe participants rated the last task only, that is the GMAIL registration.

6.3 Impact on Password Strength

We evaluate the null hypothesis $\mathsf{H}_{\mathsf{P},0}$: *Solving a CAPTCHA does not impact password strength* across both password strength measures.

6.3.1 Passwordmeter

The distribution of the Passwordmeter password strength score is measured on interval level and is not significantly different from a normal distribution for each condition. Saphiro-Wilk for (a) CONT, $D(22) = .976$, $p = .848 > .05$, (b) text-CAPTCHA, $D(22) = .967$, $p = .641 > .05$, (c) picture-CAPTCHA, $D(22) = .962$, $p = .534 > .05$. Levene's test for the homogeneity of variances show that the variances were not significantly unequal across conditions, $F(2, 63) = 0.638$, $p = .532 > .05$.

We computed an one-way ANOVA with the password strength score as dependent variable. There was a statistically significant effect of the experiment condition on the password strength score, $F(2, 63) = 6.716$, $p = .002 < .05$. We measure the effect size in Cohen's $f = .42$ from ($\eta^2 = .176$ $[0.043, 0.296]$) and Cohen's $\omega^2 = 0.148$. This constitutes a large effect. We provide the descriptive statistics in Table 1 and the means/interval plot in Figure 5. We reject the null hypothesis $\mathsf{H}_{\mathsf{P},0}$.

As post-hoc test, we conducted a Tukey HSD reporting that the password strength was statistically significantly lower in the text-CAPTCHA condition ($M = 31.05$, $SD = 29.16$) than in the control condition ($M = 67.68$, $SD = 37.02$) with $p = .002 < .05$. We have an effect size in Hedges' $g = 1.08$ $[0.44, 1.71]$.

Furthermore, the password strength in the picture-CAPTCHA condition ($M = 42.36$, $SD = 35.18$) was statistically significantly lower than in the control condition, $p = .042 < .05$. That is at an effect size in Hedges' $g = 0.69$ $[0.08, 1.29]$.

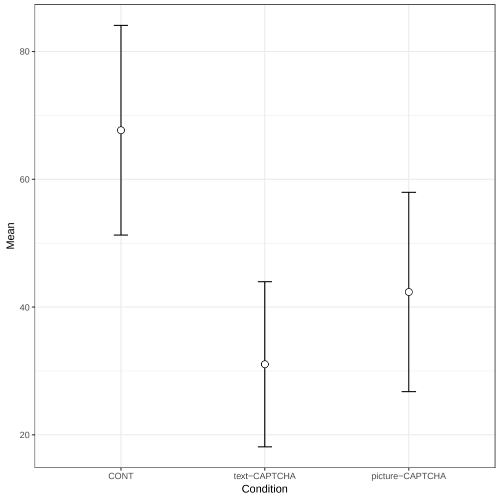

Figure 5: Passwordmeter

Figure 6: 95% Confidence Intervals on means of password strength scores by condition. (Main Experiment)

6.3.2 Zxcvbn

The distribution of the zxcvbn log_{10} guesses is measured on interval level and is not significantly different from a normal distribution for each condition. Saphiro-Wilk for (a) CONT: $D(22) = .184$, $p = .053 > .05$, (b) text-CAPTCHA: $D(22) = .148$, $p = .148 > .05$, (c) picture-CAPTCHA: $D(22) = .121$, $p = .538 > .05$. We also computed Levene's test for the homogeneity of variances. For the zxcvbn log_{10}, the variances were not significantly unequal: for CAPTCHA and control conditions, $F(2,63) = 1.072$, $p = .349 > .05$.

We computed a one-way ANOVA with the zxcvbn log_{10} guesses as dependent variable. There was a statistically significant effect of the experiment condition on the zxcvbn log_{10} guesses, $F(2,63) = 4.665$, $p = .013 < .05$. We measure the effect size in Cohen's $f = .36$ from ($\eta^2 = .130$ $[0.016, 0.244]$) and Cohen's $\omega^2 = 0.1$. This constitutes a medium to large effect size. We provide the descriptive statistics in Table 2. Based on zxcvbn, we would equally reject the null hypothesis $\mathsf{H}_{\mathsf{P},0}$.

As a post-hoc test, we computed Tukey HSD, reporting that password strength was statistically significantly lower in the text-CAPTCHA condition ($M = 6.38$, $SD = 2.84$) than in the control condition ($M = 8.66$, $SD = 2.11$), $p = .010 < .05$. We have an effect size measured in Hedges $g = 0.89$ $[0.27, 1.51]$. There was no significant difference between the picture-based CAPTCHA condition and the control condition.

Because the normality of the data was borderline, we computed a Kruskal-Wallis test on the zxcvbn log_{10} guesses as well, which showed that there was a statistically significant difference in the zxcvbn log_{10} guesses between the different conditions $\chi^2(2) = 10.340$, $p = .006 < .05$, with a mean rank zxcvbn score of 42.55 for CONT, of 23.95 for text-CAPTCHA and of 34.00 for picture-CAPTCHA.

In addition, zxcvbn also provides an ordinal score ranging from 0 to 4. We computed a Kruskal-Wallis test which showed that there was a statistically significant difference in the zxcvbn score between the different conditions $\chi^2(2) = 9.251$, $p = .010 < .05$, with a mean rank zxcvbn score of 40.86 for CONT, of 24.27 for text-CAPTCHA and of 35.36 for picture-CAPTCHA.

7 Discussion

Our findings that solving a CAPTCHA prior to choosing a password impacts the password strength, has wide implications because of the impact on authentication security. We note that the more effortful text-CAPTCHA led to weaker passwords. However although the effort spent (via Diff_Tiredness) was not significantly different between the picture-CAPTCHA and the control, there was still a difference in password strength between the two conditions.

While Groß et al.'s [12] showed that a combination of tasks specifically designed in psychology to cognitively deplete users (the white bear, an impulse control and the Stroop test), resulted in users choosing weak passwords, this research shows that even common security measures such as the CAPTCHA challenge has a detrimental effect on password strength.

In addition, while system designers often create account registration forms (such as Facebook, Reddit, Wikipedia) with a CAPTCHA challenge *after* password choice rather than before, our research informs future design decisions of the positioning of CAPTCHAs. Our findings indicate that we should clearly guide the sequence of user input for usability and not to put the password strength at risk. Design recommendations such as positioning the CAPTCHA on a separate page after password choice is likely to be beneficial for security. We also observe that CAPTCHAs are often deployed as a gateway to access Web sites at all, either when frequent requests from the originating IP address were observed or when the use of the TOR Anonymizer was detected. Consequently, in these cases users are systematically exposed to CAPTCHAs before they could register and choose a password.

Furthermore, apart from considering individual and subsequent effects on a security task, it is also important to consider the overall, combined cognitive ef-

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

Condition	N	Mean	Std. Dev.	Std. Error	95% CI		Min	Max
					LL	UL		
CONT	22	67.68	37.02	7.89	51.27	84.09	-8	151
text-CAPTCHA	22	31.05	29.16	6.21	18.12	43.97	-16	103
picture-CAPTCHA	22	42.36	35.18	7.50	26.76	57.96	-17	102
Total	66	47.03	36.82	4.50	37.98	56.08	-17	151

Table 2: Descriptive statistics of password strength via zxcvbn $\log_{10}$ guesses by condition.

Condition	N	Mean	Std. Dev.	Std. Error	95% CI		Min	Max
					LL	UL		
CONT	22	8.66	2.11	0.45	7.72	9.59	4.77	14.97
text-CAPTCHA	22	6.38	2.84	0.61	5.12	7.64	0.95	13.41
picture-CAPTCHA	22	7.76	2.46	0.52	6.67	8.85	2.86	13.34
Total	66	7.60	2.62	0.32	6.95	8.24	0.95	14.97

fort of different tasks. In particular whether they lead to the user rejecting security overall. So far past research has only looked at security fatigue of the current task [11, 24]. Therefore, our findings raise several questions (a) How does designing security tasks in sequence impact (i) usability, (ii) rejection of security and security fatigue, and consequently (iii) the overall security achieved? (b) Does a sequence of security tasks induce a weak link? (c) What combined effort can the user bear? (d) What combination of security tasks is within the user's cognitive effort capacity?

7.1 Ethics

We followed the ethical guidelines at our University to run both the pre-study and the main study. We did not induce more effort than is reasonable in daily life. The participants were informed of the approximate length of the studies, were guided through a consent form and were informed that they could cease participation at any time. Participants were rewarded with a compensation of $6.5 for their time.

Participants' data are kept securely under lock and key and on machines with hard disk encryption. The personal identifiable data of participants was separated from the experiment data and the experiment data anonymized.

7.2 Limitations

Ecological Validity. Although requiring a more controlled setup, we chose lab studies as a first step because it is believed in password research that such studies offer better data quality. We used the same GMail mockup as previous studies [12] which is identical to the GMail account registration page.

Our findings pertain to the chosen manipulations. We chose the text-recognition CAPTCHA, known to be widely used and an adapted picture-recognition CAPTCHA, which often comes up from the widespread reCAPTCHA. Further experiments can be conducted on other CAPTCHA schemes and sequence in security tasks.

Sample Size and Power The study fulfills recommendations have a power of at least $1-\beta = 80\%$ against an effect size of Cohen's $f = 0.5$ in the omnibus ANOVA. With the given sample size of $N = 66$, an effect of Cohen's $f = 0.4$ could still be detected at 80% power. We note that the ANOVA on zxcvbn slightly fell below that mark.

Given the sample size investigated, the parameter estimation on the means and effect sizes in differences is not especially tight. Further research with larger samples could tighten the confidence intervals on the population parameters.

Sampling Bias. Our sample was from university student population, hence an educated sample with a mean age of 21.79 in the main study. Although we found that password choice was weaker in text-CAPTCHA than in the control condition for both the pre-study and the main study, for generalizability, this study can easily be reproduced on a stratified sample. A larger sample size can also support other statistical analyses such as regressions.

Post Questionnaires. In the post-stress and the cognitive workload questionnaires, we found that participants evaluated the GMail registration only rather than the CAPTCHA and registration. Future experiments evaluating stress and workload combined across a sequence of tasks would benefit from making clearer to participants about the task being evaluated. It might also be beneficial to consider a small stress and workload evaluation in between the sequential tasks.

Our findings of security fatigue are focused on the security tasks chosen and the sequence in which they were designed. For example it might be useful to also compare the impact of setting a password on performance at solving a CAPTCHA or other security tasks.

We did not include subjective evaluation of the CAPTCHAs. Future studies can also benefit from additional measures such as self-report/subjective participant feedback on solving the CAPTCHA and also from the combination of CAPTCHA and password.

8 Conclusions

We provide a first empirical study evaluating security fatigue in relation to sequential security tasks. We find that password choice following a CAPTCHA lead to poorer passwords than without the CAPTCHA. While our findings impact design practice and research on individual security tasks together with their pairing with other tasks, they also have wide implications for the overall security of systems.

References

[1] A. Beautement, M. A. Sasse, and M. Wonham. The compliance budget: managing security behaviour in organisations. In *Proceedings of the 2008 workshop on New security paradigms*, pages 47–58. ACM, 2009.

[2] E. Bursztein, S. Bethard, C. Fabry, J. C. Mitchell, and D. Jurafsky. How good are humans at solving captchas? a large scale evaluation. In *Security and Privacy (SP), 2010 IEEE Symposium on*, pages 399–413. IEEE, 2010.

[3] E. Bursztein, M. Martin, and J. Mitchell. Text-based captcha strengths and weaknesses. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 125–138. ACM, 2011.

[4] K. Chellapilla, K. Larson, P. Simard, and M. Czerwinski. Designing human friendly human interaction proofs (hips). In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 711–720. ACM, 2005.

[5] M. Chew and H. S. Baird. Baffletext: A human interactive proof. In *Electronic Imaging 2003*, pages 305–316. International Society for Optics and Photonics, 2003.

[6] K. P. Coopamootoo and T. Groß. Evidence-based methods for privacy and identity management. In *Privacy and Identity Management. Facing up to Next Steps*, pages 105–121. Springer, 2016.

[7] A. S. El Ahmad, J. Yan, and L. Marshall. The robustness of a new captcha. In *Proceedings of the Third European Workshop on System Security*, pages 36–41. ACM, 2010.

[8] J. Elson, J. R. Douceur, J. Howell, and J. Saul. Asirra: a captcha that exploits interest-aligned manual image categorization. In *ACM Conference on Computer and Communications Security*, volume 7, pages 366–374. Citeseer, 2007.

[9] C. A. Fidas, A. G. Voyiatzis, and N. M. Avouris. On the necessity of user-friendly captcha. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2623–2626. ACM, 2011.

[10] D. Florêncio, C. Herley, and P. C. Van Oorschot. Password portfolios and the finite-effort user: Sustainably managing large numbers of accounts. In *Usenix Security*, pages 575–590, 2014.

[11] S. Furnell and K.-L. Thomson. Recognising and addressing 'security fatigue'. *Computer Fraud & Security*, 2009(11):7–11, 2009.

[12] T. Groß, K. Coopamootoo, and A. Al-Jabri. Effect of cognitive depletion on password choice. *Learning from Authoritative Security Experiment Results (LASER'16)(July 2016), S. Peisert, Ed*, 2016.

[13] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183, 1988.

[14] E. H. Hess and J. M. Polt. Pupil size in relation to mental activity during simple problem-solving. *Science*, 143(3611):1190–1192, 1964.

[15] P. Hoonakker, N. Bornoe, and P. Carayon. Password authentication from a human factors perspective. In *Proc. Human Factors and Ergonomics Society Annual Meeting*, volume 53, pages 459–463. SAGE Publications, 2009.

[16] D. Kahneman. *Thinking fast and slow*. Farrar, Strauss, 2011.

[17] V. Kothari, J. Blythe, S. W. Smith, and R. Koppel. Measuring the security impacts of password policies using cognitive behavioral agent-based modeling. In *Proceedings of the 2015 Symposium and Bootcamp on the Science of Security*, page 13. ACM, 2015.

[18] G. Mori and J. Malik. Recognizing objects in adversarial clutter: Breaking a visual captcha. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2003.

[19] G. Reynaga, S. Chiasson, and P. C. van Oorschot. Exploring the usability of captchas on smartphones: Comparisons and recommendations. In *NDSS Workshop on Usable Security USEC*, 2015.

[20] A. Rogers and G. Brewer. Statistics for websites using captcha technologies, 2017.

[21] S. Sivakorn, J. Polakis, and A. D. Keromytis. I'm not a human: Breaking the google recaptcha. *Black Hat,(i)*, pages 1–12, 2016.

[22] C. D. Spielberger, R. L. Gorsuch, and R. E. Lushene. Manual for the state-trait anxiety inventory. 1970.

[23] C. D. Spielberger, R. L. Gorsuch, R. E. Lushene, P. Vagg, and G. Jacobs. Stai manual for the state-trait anxiety inventory. palo alto, 1970.

[24] B. Stanton, M. F. Theofanos, S. S. Prettyman, and S. Furman. Security fatigue. *IT Professional*, 18(5):26–32, 2016.

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

[26] L. Von Ahn, M. Blum, N. J. Hopper, and J. Langford. Captcha: Using hard ai problems for security. In *International Conference on the Theory and Applications of Cryptographic Techniques*, pages 294–311. Springer, 2003.

[27] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008.

[28] D. L. Wheeler. zxcvbn: Low-budget password strength estimation. In *Proc. USENIX Security*, 2016.

[29] J. Yan and A. S. El Ahmad. Usability of captchas or usability issues in captcha design. In *Proceedings of the 4th symposium on Usable privacy and security*, pages 44–52. ACM, 2008.

[30] J. Yan and A. S. El Ahmad. Captcha security: A case study. *IEEE Security & Privacy*, 7(4), 2009.

A Appendix

A.1 Password Re-Use

We asked participants whether they registered the account via a password they currently use for any services. In the control condition as well as in the picture-CAPTCHA condition, 22.7% of the participants re-used an existing password. 36.3% of the participants re-used an existing password in the text-CAPTCHA condition. This difference is not statistically significant. While a social media password was most commonly used, none of the participants re-used a banking or retail password. Table 3 provides a detailed view of password type re-used.

Table 3: Re-Use Context

Count	Service
1	email
4	social-media
2	education
2	mobile
1	social media/mobile/education
2	social media/email/mobile/education

A.2 Password Strategies

We coded participants' password strategies. Figure 7 depicts the strategies across conditions while Sections A.2.1 to A.2.6 provide the qualitative details. There was no significant difference across groups.

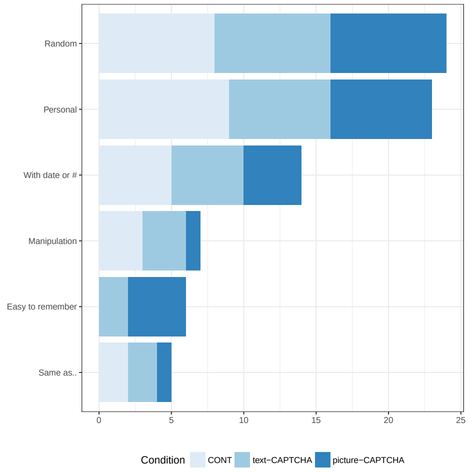

Figure 7: Password strategy across conditions

A.2.1 Random

36% of the participants did not have a strategy, described it as random or specifically said that they used a random thought.

8 participants stated that they did not have a strategy, for example, P7 expressed *"I did not have one"* or P21 in *"nothing really"*. 9 participants used a random password, such as expressed by P6 in *"I put random information together"*, by P56 in *"Trying to be funny"* or by P62 in *"random words and letters"*. 7 participants expressed a random thought, such as by P20 in *"Whatever comes to mind"*, by P29 in *"What comes to mind with different signs"* or by P58 in *"Randomly Thought of* [sic]*"*. We found that 8 participants in each condition used a random strategy.

A.2.2 Personal

36% of the participants chose a password related to their preference or something personal to them.

10 participants chose a password linked with their preferences, for example P22 expressed *"Favourite Sport with mix of capslock inbetween* [sic]*"*, or P22 *"Favourite Football player"* whereas 15 participant created a password with personal meaning, such as expressed by P26 *"Daddy's name + random letter* [sic]*"*, P30 expressed *"City where i was born.*[sic]*"* or P40 in *"My Dog's full name and the year we got him"*. We found that 9 participants created a password from a preference or with personal meaning in the control, 7 in the text-CAPTCHA and 8 picture-CAPTCHA, where 1 participant's strategy included both a preference and something personal.

A.2.3 Manipulation

We found that only 7 participants had a strategy involving complexity combinations, changing characters to numbers or the equivalent in another language, for example as expressed by P13 *"make a strong password with capital letters, small letters and numbers"* or P48 in *"a bit creative with changing the i to 1"*. 3 participants with this strategy were from the control condition, 3 in the text-CAPTCHA condition and 1 in the picture-CAPTCHA.

A.2.4 Same as . . .

Only 4 participants described a re-use strategy, for example as expressed by P3 *"Same as always"* or P47 *"Same as Username"*. We found that there was 2 participants employing this strategy in the control and 1 in each of the text and picture-CAPTCHA conditions.

A.2.5 Easy to remember

Only 6 participants reported that they created an easy to remember password, for example as expressed by P4 *"just used two easy words without spaces between which is easy to remember for me"* or P15 *"making the password unreal and easy to remember"*. 2 participants in the text and 4 in picture-CAPTCHA reported this strategy compared to none in the control condition.

A.2.6 With date or number

21% of participants created a password that was combined with numbers or dates, for example P18 reported using *"My Initials and current year"* and P27 *"Favourite colour and 100"*. 5 participants employed this strategy in both the control and text-CAPTCHA conditions and 4 in the picture-CAPTCHA condition.