Finding Secret Treasure? Improving Memorized Secrets Through Gamification

Katrin Hartwig
Atlas Englisch
Jan Pelle Thomson
Christian Reuter
hartwig@peasec.tu-darmstadt.de
me@lu-e.de
me@skeleten.me
reuter@peasec.tu-darmstadt.de
Science and Technology for Peace and Security (PEASEC), Technische Universität Darmstadt
Darmstadt, Germany

ABSTRACT

Users tend to bypass systems that are designed to increase their personal security and privacy while limiting their perceived freedom. Nudges present a possible solution to this problem, offering security benefits without taking away perceived freedom. We have identified a lack of research comparing concrete implementations of nudging concepts in an emulated real-world scenario to assess their relative value as a nudge. Comparing multiple nudging implementations in an emulated real-world scenario including a novel avatar nudge with gamification elements, this publication discusses the advantages of nudging for stronger user-created passwords regarding efficacy, usability, and memorability. We investigated the effect of gamification in nudges, performing two studies ($N_1 = 16, N_2 = 1,000$) to refine and evaluate implementations of current and novel nudging concepts. Our research found a gamified nudge, which integrates a personalizable avatar guide into the registration process, to perform less effectively than state-of-the-art nudges, independently of participants' gaming frequency.

CCS CONCEPTS

• **Security and privacy → Social aspects of security and privacy**.

KEYWORDS

nudging, cybersecurity, usable security

ACM Reference Format:
Katrin Hartwig, Atlas Englisch, Jan Pelle Thomson, and Christian Reuter. 2021. Finding Secret Treasure? Improving Memorized Secrets Through Gamification. In *European Symposium on Usable Security 2021 (EuroUSEC '21), October 11–12, 2021, Karlsruhe, Germany*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3481357.3481509

EuroUSEC'21, October 11–12, 2021,

ACM ISBN 978-1-4503-8423-0/21/10…$15.00
https://doi.org/10.1145/3481357.3481509

1 INTRODUCTION

Despite the existence of various strong systems for user authentication like hardware tokens, one-time passwords or password management tools, users still prefer using passwords as their main authentication factor [36, 54]. Unlike physical authentication devices, passwords – as plain text or in derived formats such as hashes – may be stolen in great quantities [8], for instance to be used in illegal data or identity theft schemes. Even if stolen, login data is protected by hashing, although even a hash will always be guessed ("cracked") given infinite time or computational resources. Therefore we must increase demands on either time or resources required to break a hash.

One option to increase the time needed to crack a password is by increasing its length and complexity so that it is not feasible to brute-force the password, and it is less susceptible to dictionary-based attacks. Unfortunately, users tend to choose passwords of weaker complexity or misjudge short, but complex and seemingly random character compositions as suitable and secure passwords [9, 15]. To mitigate these risks, current research is evaluating *nudging* concepts to steer users towards stronger password choices. A nudge *"alters people's behavior in a predictable way without significantly changing their economic incentives"* [47, p. 6]. After being successfully established in contexts such as the health sector, it has more recently become a research topic in cybersecurity as well [1].

This paper heavily focuses on the user as the decisive factor in the chain of information security. As Garfinkel and Lipford [16, p. 8] state, *"only by simultaneously addressing both usability and security concerns will we be able to build systems that are truly secure."* In other words, users will always find ways to ignore or disable security measures forced on them – e.g., for login processes this could be achieved by slightly altering their universally used password to fit the system's complexity requirements, with diminishing returns. We must instead strive to include users in the process of securing their data in a way that does not overwhelm or hinder them. Nudging is a promising concept of assisting users in adopting a strong password policy without forcing it on them.

Current guidelines [5, 18] advise service providers not to impose restrictions on the composition of passwords, except for limiting the password choices to those of eight characters or longer and possibly checking against a banlist of common passwords. This puts

the responsibility of choosing a strong password solely on the user. Thus providing users with nudges to increase password strength without altering the requirements presented by the service provider could lead to greater security. At the present time, research is ongoing regarding the efficacy of such nudges and ideal nudge selection [20]. To contribute to existing research, we compared current nudging approaches with novel concepts to increase password security and provide transparent help cues to nudgees. Some approaches include elements of gamification to make nudges more effective and approachable by increasing the user's motivation to change their learned behavior [29, 31]. According to Karimi and Nickpayam [26, p. 34], gamification describes *"a strategy that employs game mechanics, techniques, and theory in areas that traditionally do not function like a game."* Silic and Lowry [45] for instance applied this concept to the context of cybersecurity.

We expected gamified nudges to resonate more strongly with audiences used to playing digital games than nudges that do not provide gamification elements, and aimed for users to accept these as a less authoritative nudging option compared to strict policies and forced restrictions. Previous research [37] has highlighted individualized nudges to be more effective than a one-size-fits-all approach. Hence, contributing to the trend of personalization, we evaluated if gamified nudges were more suitable for audiences that play digital games while providing different nudging options, which may then be selected based on the target audience of the provided service. We avoided approaches that significantly lengthen the registration process, as well as nudges that are severely intrusive as not to frustrate users [38, 53]. In conclusion, we address the following research questions:

(Q1) How do gamified nudges compare to a state-of-the-art password meter?

(Q2) Does a user's gaming frequency increase the acceptance of a gamified nudge?

The paper is structured as followed: First, we present the body of related literature resulting in a research gap (see section 2). Subsequently, we describe our research design (see section 3). In section 4, we present the study design and results of our preliminary think-aloud study, including a detailed description of all investigated nudges, followed by section 5 where we present method and results of our main study. We complement our work by a discussion of results (see section 6), entailing a presentation of limitations of our work and relevant implications for the future, followed by the conclusion in section 7.

2 RELATED WORK AND RESEARCH GAP

We can identify several current trends within the research field of nudging in cybersecurity. In the following we will focus especially on two aspects: gamification and personalization. Some works have highlighted the potential of enhancing motivation to engage in a more secure behavior through gamification. Other studies have brought up the idea of using personalized nudges instead of one-size-fits-all nudges to be more effective.

2.1 Gamified Nudges in Cybersecurity

While there are several studies dealing with nudging in cybersecurity, researchers have pointed out that nudges are often not as effective as desired in this case. For example, Kankane [25] evaluated the effectiveness of different nudges for password management. They found that none of the nudges was effective enough to significantly change the individuals' behavior regarding password creation. Against this background, the concept of gamification has more recently emerged as a potentially promising trend in cybersecurity to encourage engagement and intrinsic motivation and, thus, might make nudges more effective. When looking at the approaches in detail, we can differentiate between user interventions that use gamification as preventive training and gamified user interventions that take place at the point of a critical decision, e.g. as a nudge when about to create a password.

On the training interventions, Silic and Lowry [45, p. 5] conducted a long-term field study with 420 employees within an organization. As the gamified user intervention, they developed and evaluated a gamified website including an avatar and different game mechanisms such as the possibility to earn points after completing quizzes and reading tips about security education topics. Hence, the applied user intervention focuses strongly on educational elements that take place as preventive training. They found that gamified educational user interventions in form of security training can indeed have a significantly positive effect on security behavior, fulfilling users' motivations and coping needs [45]. The idea of educational games in cybersecurity is not completely new, as Sheng et al. [44] designed and evaluated Anti-Phishing Phil in 2007, an educational online game that teaches users not to fall for phishing attacks. The authors compared several training conditions (e.g., reading anti-phishing tutorials, reading online training material or playing the Anti-Phishing Phil game) and found that participants performed better in identifying fraudulent websites when being assigned to the game condition beforehand. Canova et al. [6] followed up on that idea by developing NoPhish, an anti-phishing education app to teach users accessing, parsing and checking URLs regarding potential phishing attacks, raising the security awareness. Focusing on the context of password security awareness, Scholefield and Sheperd [43] developed a mobile-based application with gamification features. Basis of this application is a multiple-choice quiz, educating users on topics such as choosing a strong password and avoiding common passwords. Overall, the application received positive feedback from participants of a small pilot study ($N = 17$).

Less attention has been paid to approaches that integrate gamification elements as a user intervention at the critical point of decision. For instance, Takada and Hattori [46] investigated how users can be motivated to voluntarily use a secure pattern-based authentication when provided with a role-playing game function to the authentication process. The authors state that while there are methods such as password policies and password meters to assist users in using secure credentials, those measures often come with drawbacks in usability and user experience. In a preliminary online experiment with eight participants, they found that a role-playing game function within the authentication process does indeed have the potential to motivate users to use stronger credentials [46]. Also, Ur et al. [49] evaluated if a dancing bunny as a gamified password meter led to longer passwords, resulting in better efficacy but mixed qualitative feedback. Further, Micallef and Arachchilage [32] designed and evaluated a gamified nudge to improve users' memorability of security questions. They adapted the 4 Pics 1 Word mobile

game, and found that a gamified nudge has the potential to improve the memorability of answers to security questions, constituting a promising approach to reduce the trade-off between usability and security in fall-back authentication [32]. Ophoff and Dietz [35] performed an online experiment with 232 participants to compare the effectiveness of password strength feedback with and without gamification elements. Therefore, the participants interacted either with a common password meter or a gamified feedback method where they could earn points while completing the authentication process. Different from our experiment, they chose a simple bar indicating strength by color (e.g., green to red), length of the bar, and text (e.g., "strong") instead of the state-of-the-art password meter by Ur et al. [50]. The gamified feedback dynamically displays a score of password strength for each character that was added or removed, showing the difference of strength to the previous password. When comparing the results of the test conditions in the online experiment, password cracking time did significantly differ between participants that were assigned to the password meter and those that interacted with the gamified nudge, showing potential for the gamified feedback [35].

2.2 Personalized Nudges

Currently, nudges are mostly implemented as a one-size-fits-all solution. However, to include users in the process of securing their data in a way that does not overwhelm or hinder them, personalization of nudges seems to be a promising approach. It focuses on adapting the design to individual users' requirements to make nudges in cybersecurity more effective and beneficial to a diverse set of end users instead of the average user alone [11]. Various studies have already investigated this idea or suggested potential benefits in their works [11, 22, 28, 37, 39, 48]. Some argue that such an approach might improve compliance significantly [11]. For instance, Knijnenburg [28] points out that showing tailored nudges can support users in making better privacy decisions.

To provide end users with the subjectively most effective nudge, it is fundamentally necessary to identify distinct user groups. Hartwig and Reuter explored people's attitude towards nudging in cybersecurity for different contexts to gain an understanding of how to address certain users [21]. More practically, various approaches use psychometric scales to segment users. For instance, Egelman et al. [11] use decision-making styles and risk-taking attitudes to predict privacy and security behavior and segment user groups accordingly. This follows the assumption that understanding users' attitudes towards computer security helps to contextualize their observed behaviors as well as with predicting their future behaviors. Further, Dupree et al. [10, p. 5228] were able to identify distinct categories of end users by analyzing the participants' attitudes and behaviors towards security practices. Based on these findings they suggest utilizing these clusters in the design of new privacy and security tools.

Even though the personalization of nudges is considered promising by several studies, only a few have implemented the concept within the cybersecurity context. Pe'er et al. [37] tested people's decision styles to personalize nudges for stronger passwords in two online experiments ($N = 2,047$) and argue that choosing a nudge from a pool of multiple existing nudges could be more effective than showing the same nudge to everyone. They achieved stronger passwords with personalization than with one-size-fits-all nudges and showed that decision-making styles can indeed be used to personalize nudges [37]. Research from Jeske et al. [22] also suggests that user differences play a role in security decision-making. They point out that the effectiveness of nudges depends on user characteristics, such as their impulse control when selecting a public wireless network. Hence, the personalization of nudges in cybersecurity seems to be a more promising approach than one-size-fits-all nudging.

To allow many individuals to benefit, it is necessary to create a pool of different nudges. For instance, gamification was identified as a promising concept to motivate secure behavior and to encourage engagement and intrinsic motivation. Hence, we consider it worthwhile to investigate the effect of integrating elements of gamification in nudging and to shed light on the question if users with different gaming preferences favor different types of nudges.

2.3 Research Gap

Nudging in cybersecurity has recently emerged into a growing research field, resulting in the design and evaluation of different nudges for that specific context (e.g. [1, 20, 22, 28, 32, 37, 41, 49]). Yet, researchers have found that nudges in cybersecurity are often not as effective as desired [25]. Some argue that personalization instead of one-size-fits-all nudges is a promising trend to enhance effectiveness (e.g. [28]). Also, using elements of gamification in nudges can be considered a potentially effective measure to increase motivation of end users to behave in a more secure way. The benefits of gamification have already shown in the context of cybersecurity. Micallef and Arachchilage [32] have investigated the potential of gamified nudges in the context of fall-back authentication, potentially reducing the trade-off between usability and security. Further, Ophoff and Dietz [35] have made initial investigations to evaluate gamified password feedback in comparison to a simple password meter. However, most current user interventions using gamified elements in cybersecurity focus on preventive training instead of nudging at a point of critical decision-making. The context of manual password creation is still highly relevant in usable security, as passwords are still the number one choice of user authentication despite various strong alternatives and end users tend to struggle following password requirements while creating usable and strong passwords. To our knowledge, password nudges with gamification elements considering gaming preferences as an indicator for personalization have not been investigated on a larger scale. We hypothesize that especially an avatar as a customizable assistant has the potential to combine benefits of gamification and personalization as motivating factors.

Therefore, we suggest taking a closer look at password nudges with gamification elements and compare their effectiveness among users with different gaming preferences. By doing so, the pool of effective nudges in cybersecurity may be extended, facilitating personalization. Therefore, we performed a two-fold evaluation of different nudges for stronger passwords, focusing on users within the German population to provide an opportunity for comparison with other countries in the future.

3 RESEARCH DESIGN

In a two-fold evaluation we first conducted a qualitative think-aloud study with seven nudge prototypes. The initial set of nudges contained both state-of-the-art nudges as well as novel nudges incorporating password feedback with gamification elements. For our subsequent main study, we reduced the sample of nudges to a novel gamified nudge that received positive feedback, a state-of-the art nudge for password strength and a state-of-the-art nudge for memorability. We refined these three nudges, considering feedback we obtained from our participants during the preliminary study. In our main study we conducted an online experiment with 1,000 participants, testing the effect of the three different and promising nudges on password strength, short-term memorability, and usability in comparison to a control group without a nudge. Both studies were conducted in accordance with the requirements of the local ethics committee of the TU Darmstadt.

In the following, we present important background information on how we calculated password strength, and which password criteria were considered for our nudges. An important resource for password creation rule sets is the 2016 NIST report [18], which states that the only rules to be forced on users may be a minimum password length of eight characters and that the password may not be included in an optional ban list, e.g. containing previously leaked passwords. To assist users in creating stronger passwords we, however, consider it sensible to give optional suggestions for possible password compositions (e.g., including digits). In accordance with the NIST report we decided to give those suggestions not as a forced password criterion but as a suggestion to encourage variety. Slightly modifying the idea of Komanduri et al. [30], we have decided on a uniform standard for password criteria and suggestions in our nudges. As password criteria, we requested at least eight characters as suggested by NIST. We further checked against a ban list of 100,000 common passwords, included in the service library NBP. For optional suggestions, we proposed including two or more characters per character class (upper case letters, lower case letters, digits, special characters). Additionally, we have included an informative text regarding the creation and criteria of strong passwords, featuring an example of "mnemonics", a memorable password strategy following work by Kiesel et al. [27]. As Kaleta et al. [24] suggest, we have included a section explaining why secure password choices are important and have an impact on information security (see Figure 3 in the appendix).

4 PRELIMINARY THINK-ALOUD STUDY

We conducted a preliminary study ($N = 16$) in individual think-aloud sessions to qualitatively investigate the comprehensibility and usability of several different password nudges. The participants were acquired through Twitter and Discord channels (e.g., the university channel) and were between 20 and 30 years old students, mostly of a technical degree program. The sample strategy has to be seen in the context of our two-fold study design. While the participants in the preliminary study were selected from the investigators' direct environment without aiming at a systematic sampling, the subsequent large-scaled online experiment of our main study is based on a broader demography, and validates and expands preliminary findings of the think-aloud study quantitatively.

Table 1: Initial set of nudges. Nudges with the * asterisk were refined for the main study.

Key	Description	Source
N1*	Dynamic meter with checklist	[50]
N2*	Generated default password	[2, 7]
N3	Radar chart	[20]
N4	Expectation and Reflection	[7, 40]
N5	Reminding of consequences	[7, 19]
N6	Interactive chat	*novel*
N7*	Avatar creation	*novel*

Comprehensibility and usability are important factors for password nudges to be effective. Therefore, conducting a qualitative study as a first step helps us to identify potentially promising directions for novel nudges in comparison to established state-of-the-art nudges. The think-aloud study further helped us to refine promising nudges and to choose a smaller set of nudges for our large-scale main study. In the following, we present how the initial set of seven nudges was chosen and how the evaluation was conducted. We follow up with the presentation of the results of our preliminary study, motivating the research design of our subsequent main study.

4.1 Nudge Selection

For the preliminary study, we have created seven prototypes (see Figure 1), of which three have been refined for use in the quantitative study (marked with an * asterisk in Table 1). Based on existing literature on password nudges, for the initial set of nudges we made sure to include state-of-the-art nudges for password strength and memorability to have a valid basis for comparison. Therefore, the dynamic meter with an integrated checklist by Ur et al. [50] was included as N1, representing the state-of-the-art for strength enhancement. We further included a generator of default passwords (N2) as a state-of-the-art nudge for memorability, based on the ideas of Caraban et al. [7] and Al-Ameen et al. [2]. As we aim to compare gamified with non-gamified nudges, we additionally included promising nudges that take advantage of different psychological effects such as reminding of the consequences and reflection without using gamification as N4 and N5, following the suggestions of Caraban et al. [7]. As gamified nudges, we included a radar chart with potential gamification elements for password feedback (N3) that has already been investigated by Hartwig and Reuter [20] to some extent, missing information on memorability. Finally, we included two novel nudges with gamification elements, namely an interactive chat and the creation of an avatar as a customizable assistant (N6 and N7). In the following, we describe each nudge in more detail.

N1*: Dynamic Password Meter. The dynamic password meter with a checklist is currently one of the most successful approaches to password nudging, albeit competing with a more simple password meter which is commonly used in registration forms. In our prototype, both are featured: a data-driven password meter making use of a traffic light color scheme, and a checklist detailing which required and suggested password criteria are met (see Figure 1 at the top left), similar to the work of Ur et al. [50]. The nudge combines

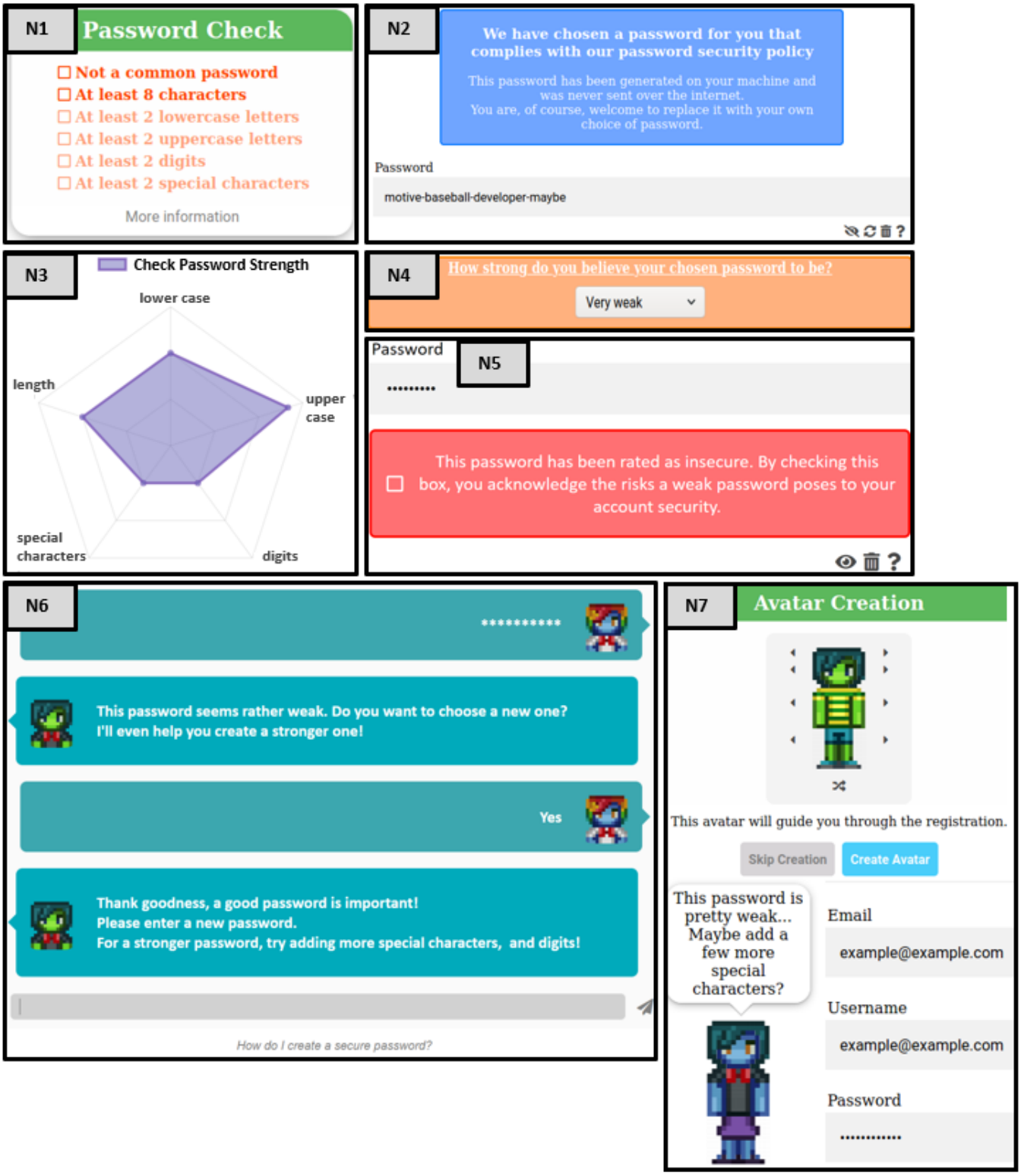

Figure 1: Representation of all seven investigated nudges N1 to N7. (own figure)

the "at a glance" nature of a meter with the transparent approach of a checklist. Data-driven password meters in the style of Ur et al. [50] are regarded as the state-of-the-art concept in nudging and, modified to fit the password criteria specified, has been our baseline for testing the efficacy of novel nudges.

N2*: Generated default password. Caraban et al. [7] discussed the concept of providing a default option for users, which we have implemented based on the generation method suggested by a XKCD web comic [33]. As examined in related work [2, 17], this method should provide significant memorability benefits compared to random string generation, making it our baseline for testing the short-term memorability of our nudges. The user is presented with a default password generated on the client, which intends to provide a secure base choice, suggesting what a secure password may look like and encouraging the use of a service-unique, random password (see Figure 1 at the top right). We examined if users tend to choose a generated or a custom password. The user may repeatedly choose to generate a new default password to aid in finding an easy to memorize secret.

N3: Radar chart. The radar chart, as described by Hartwig and Reuter [20], aims to provide a transparent view into how a passwords' strength is measured and how a user may improve their password. Along every axis radiating from the center, a criterion of the total password strength is quantified (see Figure 1 at the middle left). As this is a chart, reading off and understanding how well the criteria are satisfied should feel more intuitive to the nudgee than an aggregated score. This approach features elements of gamification, which could allow us to compare our novel gamification-inspired nudges to one already discussed in the present literature.

N4: Expectation and Reflection. As human decision making tends to fall into two categories, driven by either fast and instinctive or slow and deliberate thinking [23], this nudge aims to make use of the more deliberate system for password selection. By providing the user with the option to rate their password's strength on a subjective scale before submitting the registration, our intent was to invoke reflection and re-evaluation regarding the security of the chosen secret. As this nudge does not provide active feedback about the password strength, we took interest in comparing it to the novel nudges, which provide explicit suggestions and feedback. Similar work to this nudge has been done by Renaud and Zimmermann [40].

N5: Reminding of consequences. Very similar in concept to the previous nudge N4, but different in how reflection is achieved, N5 checks the zxcvbn score of the password and, if rated weak, forces the user to check a box confirming they understand some of the risks and consequences of choosing a weak password as detailed in the info text (see Figure 3). Our implementation is inspired by the eponymous section in Caraban et al. [7]. We examined which of these two nudges is more effective in strengthening a password, so one may be chosen for comparison in the main study if deemed promising.

N6: Interactive chat. Some services have adopted gamification elements into their sign-up process by implementing a chat-based registration. The user chats with an automated system (a *bot*), which gives feedback on password strength. It queries the user's information by asking questions in natural language and processing the reply (see Figure 1 at the bottom left), although it usually does not understand responses in natural language (e.g. "My name is Bob."), but rather concise information (e.g. "Bob"). Chatbots have already been discussed in academic research (e.g. [3]). However, to our knowledge, the efficacy of a chat-based approach as a password nudge was not previously evaluated.

N7*: Avatar creation. Following the trend of gamification, we have introduced a digital avatar to the registration process (see Figure 1 at the bottom right). In accordance with related research (e.g. [51]), the appearance of the avatar is customizable. It intends to create a bond between the user and the avatar, with the targeted outcome of increasing the value of its suggestions and encouraging secure passwords by offering context-aware positive feedback. Following our goal of not significantly lengthening the process, the avatar runs parallel to the registration. It offers information without requiring interaction and may be skipped entirely (however not in our experiment). The hypothesized success of this concept is based on multiple psychological effects, predominantly the "IKEA effect" [34], giving weight to the avatar's cues by including a "do-it-yourself"-styled personalization section in addition to the perceived "sunk cost" of spending time customizing, which adds further incentive to satisfy the avatar's suggestions.

4.2 Methodology

We implemented high-fidelity and fully functional prototype versions of the nudges for our preliminary study. Those provided the opportunity for user interaction within a role-playing registration form. A mock registration form was implemented to provide a realistic registration process, allowing a role-playing approach for preliminary evaluation of the nudges. Previous research [12] has shown this strategy to increase the probability of study participants acting similar to real-life situations. In individual think-aloud sessions, the acceptance of various nudges was assessed and explicit feedback on implementation and concept was gained. Hence, we focused on qualitative feedback. With this data, we gained first qualitative insights, filtered out unpromising nudges and further refined the remaining approaches for use in the second, large-scale study. The participants of the preliminary study each reviewed a subset of three or four of the initial nudges in random order, sharing their reasoning and impressions while completing the mock registration processes. Each nudge was tested by eight people. We used online-conferencing tools that allowed both screen sharing as well as audio communication to create audiovisual recordings for later evaluation.

The participants were briefed on the think-aloud method [13] and agreed to the recording of their session. For each nudge, the participants were then asked to role-play creating an account for a fictional service provider. If participants fell silent for a few seconds they were reminded to keep thinking aloud. After they successfully created an account, the participants were asked several open questions (see Table 2) which were tailored to the specific insights we wanted to gain for each nudge. While we had a prepared set of questions, we gave enough space for topics beyond. Afterwards, the participants were asked to perform a memorability test while thinking aloud where they were redirected to login with the password used during registration. The user was informed if an attempt was incorrect. We transcribed the audiovisual feedback we received and

systematically collected our findings regarding user assessments for each nudge in a qualitative content analysis.

4.3 Results

The goal of the preliminary study was to gain first qualitative insights about novel and state-of-the-art nudges with and without gamification elements before conducting a subsequent large-scale online experiment with the nudges that got a largely positive feedback. We further used this as an opportunity to collect feedback on possible improvements of the nudges' presentation.

N1*: Dynamic Meter. The state-of-the-art dynamic meter nudge is one of the most common approaches to password nudging. The input of one of our participants reflects this, who described the nudge as a "known environment". Because of the common use of this nudge, the acceptance of this nudge seemed quite high. Even though one participant criticized that the checklist seems to "apply pressure", most of the participants felt like the nudge improved their password, and described the tool as "intuitive". Therefore, it served as our baseline for testing the efficacy of novel nudges in the subsequent online experiment. For the main study we improved the distinction between mandatory and optional requirements as well as the distinction between password meter and checklist.

N2*: Generated default password. Compared to random string generation, this method should offer significant memorability benefits. Therefore, this nudge serves as our baseline for testing the short-term memorability for our nudges. We received some negative feedback on that nudge. For instance, participants were concerned about dictionary attacks when using default passwords. We chose to examine this nudge in our main study although noting some participants expressed concerns, that a recommended password feels insecure and a password persisting of four words might be too long to remember.

N3: Radar chart. This nudges' aim is a comprehensible view on how password strength is measured as well as how a user may improve their password strength using a chart-based visualization. Therefore, the user is quite involved in the process. One could argue that this nudge features gamification elements. However, we received rather negative feedback on this nudge in our preliminary study. Participants described this nudge as too distracting from the actual password creation, unnecessarily complicated, non-intuitive, and overly long. Additionally, they presumed this type of chart to be rather unknown to most users. Many participants noted that they "felt urged to fill out all axes" (P #140) of the chart, disregarding the creation of a complex and memorable password, instead appending numbers and special characters until the chart was filled. One participant suggested color codes like a traffic light color scheme within the chart might be helpful, another participant stated they would like the idea of this nudge as a plug-in or as default on lots of websites. Due to the mainly negative feedback, however, we did not consider N3 for our main study.

N4: Expectation and Reflection. This nudge makes use of the more deliberate system of password selection by providing the user with the option to rate their password's strength on a subjective scale before submitting the registration. This type of nudge does not provide feedback about the password strength. A lot of our participants strongly criticized this feature as being confusing, because they did not know why they were asked to reflect on their password. They described this lack of suggestion as "unhelpful". As a consequence, it was quite hard for the participants to evaluate how to strengthen their passwords. In addition, this feature made the nudge feel unintuitive due to the lack of consequences of the selection the participant executed. We did not further evaluate it in our main study.

N5: Reminding of consequences. Similar to N4 this nudge does not provide feedback, and also relies on self-reflection. However, the nudge tests the zxcvbn-score of the password the user chose and if it is rated weak, the user has to check a box confirming they understand some risks and consequences of choosing a weak password to continue the process. Participants noted that this warning is quite interesting and helpful as well as reasonable and not being intrusive. At the same time, participants describe this warning as not being transparent. The user is left in the dark why their password is weak. Therefore, some participants are left irritated, others even mistrust the tool because of this lack of transparency. Hence, we chose not to further evaluate this nudge in our main study.

N6: Interactive chat. This nudge is a chat-based registration that uses gamification elements. Some of our participants described the avatar as "cute" and "entertaining". The feedback using natural language via the bot attracted positive attention from some. It was noted that it felt like undergoing the registration process with somebody by one's side. Thus, this nudge was described as a user-friendly tool that could be quite helpful for users which are not very computer-oriented. At the same time, participants raised awareness to some security concerns: "it felt like you have to reveal your password to someone". Additionally, it stood out negatively that the nudge did not provide enough feedback about one's password strength. Aside from that, it did not feel transparent, because it was not clear what information was needed and how long the process would take. Moreover, participants criticized that the nudge is not appropriate for users who are technically oriented, and it takes away the known workflow of creating a password. Furthermore, the participants noted different faults of the design and criticized the "unnecessarily complex design" (P #861). Even though this nudge received mixed feedback, the negative responses predominated. Hence, we chose not to further evaluate it in our main study.

N7*: Avatar creation. Similar to N6, this nudge uses gamification elements in the form of a digital avatar for the registration process. This nudge received much better feedback from the participants. Some described the avatar as "cute" and "friendly", while others described it as "childish" and suggested it might therefore not be fitting for all services. Even though the phrasing of the hints was sometimes seen as "unprofessional", most of the participant rated the hints as useful and easy to understand. After improving the dialogue of this nudge as well as the hints for the avatar creation, we used it as our gamification nudge in our main study. Furthermore, to get uniform results in the main study, we deactivated the button to skip the avatar creation.

Our preliminary study aimed to gain first insights, which will be validated and quantified in our subsequent online experiment for a reduced sample of nudges. Feedback for N3, N4, N5 and N6 was rather critical. By comparison, N1 and N7 scored better in the

preliminary study. N2 received mixed feedback, but serves as our baseline for testing the short-term memorability. Therefore, we implemented the technical feedback we received for N1, N2 and N7 and used them in the subsequent online experiment.

5 MAIN STUDY: REPRESENTATIVE EVALUATION

We conducted a large-scale ($N = 1,000$) quantitative study to determine the efficacy, acceptance, and short-term memorability of our final selection of nudges, compared to a control group that did not interact with a nudge. Our goal was to compare our approaches regarding these categories, and draw conclusions about the performance of our novel avatar nudge N7 as a user intervention with gamification elements. As the final set of nudges for our main study we used N1, N2 and N7. N1 implements a state-of-the-art nudging concept by Ur et al. [50], which includes a checklist for password criteria and a one-dimensional password meter displaying the password complexity [52], providing a baseline for efficacy in nudging for stronger passwords. It was randomly assigned to 239 participants. We further included N2 "Default Password Generator", as our memorability-heavy baseline nudge – an additional nudge without gamification elements. It presents the user with a locally generated suggested password, which should provide a secure default option as well as a good password memorability for those which picked the suggested password [17] and was randomly assigned to 256 participants. Further, we included N7, the avatar as our novel gamified nudging concept, which provides natural language cues to the user. We hypothesized that when letting the user customize their own avatar, the sunk cost would give weight to the avatar's cues and suggestions. It was randomly assigned to 258 participants. We further included a control group ($N = 247$) that did not interact with a nudge while creating a password. The main study was based on a mock-up registration for a fictional streaming platform.

5.1 Method of Data Collection and Analysis

Using the German panel provider Respondi, we conducted an online experiment and integrated survey with 1,000 participants approximately representative of the German demography regarding age from 18 to 74 years, gender and education. The survey was implemented using the software LimeSurvey, and did not introduce the participants to the topic of nudging, as not to bias them, instead presenting a focus on optimizing security in online registration processes. After performing demographic checks regarding age, gender, and education, the participants were invited to complete a mock-up online registration for a fictional streaming platform, which included either a randomly assigned nudge or the control group. Afterwards, a System Usability Scale (SUS) [4] questionnaire regarding the presented nudge was filled out by the participants. As our participants were German, we used the German SUS adaption by Ruegenhagen and Rummel [42]. It provides a "quick and dirty" method for measuring usability and acceptance of a system. The control group was not exposed to any explicit nudging concept, and was therefore also not assigned to the SUS. Additionally, all participants were asked to provide further context to the experiment's results, e.g., by stating how worthy of protection they assess login data for a streaming platform. Finally, the participants were asked to perform a short-term memorability test identical to the one used in our prestudy: The participants were prompted to login with the password used during registration and were informed if an attempt was incorrect. After a maximum of five attempts to recall the correct password, the test was closed. We tracked the number of attempts required to input the correct password.

For our main study, we collected the zxcvbn-score which approximates the magnitude of guesses (log(*guesses*)) an adversary would need to crack each respective password, and grouped them by individual nudge or control group (see Table 3 in the appendix). Additionally, we compared the capped zxcvbn-score from 1 to 4 to gain more simplified and intuitive insights. For interpretation of the results it is however crucial to note that those capped results do not consider higher scores than 4. For our *suggested password* nudge N2, we also introduced the subgroups N2.d (only participants that picked the offered *default* option) and N2.c (participants that chose to use their own, *custom* password) in the test. Further, we collected the amount of attempts to recall a password in our short-term memorability test (capped at five tries) and the SUS score for usability testing. We also included demographic data and information about gaming habits, asking about the frequency with which the participants played games (mobile, desktop or any other digital format). That was crucial to identify potentials for personalization.

For pre-processing and analysis we used Python and R. To counter the problem of multiple comparisons, the determined p-values were adjusted with Bonferroni correction. Plots and tables where then derived and created using R. We first performed analyses of variance (ANOVA) to investigate if a significant statistical relevance was present. In a post-hoc Tukey HSD test, we then tested if individual data pairs significantly differed from each other. As the data for cracking time (log(*guesses*)) does not meet the requirements of an ANOVA due to outliers, in that case we performed the non-parametric Kruskall-Wallis rank sum test and a post-hoc Wilcoxon test.

5.2 Results

In the following we will present the results of our main study including user characteristics and the comparison of nudging efficacy regarding password strength, usability and memorability. We thereby especially focus on potential effects of gaming habits.

5.2.1 User Characteristics. Even though the study is not absolutely representing the German demography it comes very close concerning age, gender, and education. There were 47.8% (German average: 49.6%) female participants whereas 52.3% (50.4%) were male. Unfortunately, our panel provider does not yet support diverse gender - this limited the amount of self-reported diverse-gendered participants to 1. Age 60-74 years constitutes the most represented group with 25.4% (26.3%), followed by the age of 18-29 with 22.6% (18.8%). Meanwhile, the age group of 50-59 years constituted 19.9% (20.7%), the age group of 40-49 years 18% (18.2%), and lastly, the age group of 30-39 years was represented with 14.6% (16.1%) of our participants. We gathered information on the highest level of education in groups (without a school diploma / certificate of secondary education ("Hauptschulabschluss"), a general certificate of secondary education ("mittlere Reife"), qualification for university entrance ("Abitur"), or university degree). We observed that

our pool of participants skewed towards users of "high" education. 42.8% held a university degree or the qualification for university entrance (33.7%). 34.6% (32.5%) held a general certificate of secondary education, 23.1% (33.9%) have a certificate of secondary education or were without a school diploma. To gain context on the experiment's results, we asked our participants on a Likert scale from 1 to 5 how worthy of protection they assess login data for a streaming platform (such as in the experiment). The mean answer was $M = 3.44$ with $SD = 1.07$. For the usability score of our nudges it made no difference, however participants that rated the login data as very worthy of protection created significantly stronger passwords (e.g., for the capped zxcvbn-score $M = 3.08, SD = 1.01$ when the answer was "5" versus $M = 2.58, SD = 1.12$ when the answer was "1").

5.2.2 Comparison of Password Strength. The Kruskall-Wallis test for log(*guesses*) showed a significant difference between the test conditions (N1, N2, N7, control group): $H(4) = 518.59, p < .001$. The Post-hoc Wilcoxon test revealed that, with the exception of N1 and N2.c ($p = .57$), all group pairings showed significant statistical differences ($p = .04$ for N7 with N2.c and $p < .001$ for all other combinations) regarding password strength. We hypothesize that N2.c performed comparably to N1 regarding password strength because the *suggested password* nudge increased the participants' motivation to create a password that they assumed to be of similar strength to the suggested option. We have found that, compared to the control group in particular, all nudge groups show a statistically significant improvement in password strength, the least difference yielded by N7 and the largest improvement yielded by N2.d (see Figure 2 left).

The ANOVA test for the capped zxcvbn-score from 1 to 4 showed significant differences as well ($F(3, 996) = 129.7, p < .001$). The post-hoc Tukey HSD test revealed significant differences between all groups. The descending order of mean scores is as followed: N2.d ($M = 4, SD = 0$), N1 ($M = 3.07, SD = 0.92$), N2.c ($M = 2.91, SD = 1.07$), N7 ($M = 2.65, SD = 0.96$) and N0 ($M = 2.27, SD = 1$). Hence, while all evaluated nudges were effective regarding password strength compared to the control group, the generated default passwords were the strongest. Also, the dynamic password meter (N1) following the idea of Ur et al. [50] yields good results. The novel gamified avatar nudge (N7) is still an effective nudge, however not as strong as the other evaluated nudges. Interestingly, participants that were assigned to the generated default nudge (N2) and decided to not use the suggested password but created their own (N2.c) yield surprisingly strong scores as well. When comparing the efficacy of our nudges separately among the user characteristics age and education for insights into potential personalization, we found no significant differences (e.g., regarding age and N7: $F(4, 253) = 0.8, p = .5$).

5.2.3 Comparison of Usability Rating. To measure the relative acceptance and usability of our test conditions, we performed a similar analysis over the System Usability Scale scores, although not adding the subgroups for N2 and leaving out the control group without a nudge, which did not perform the usability test (see Figure 2 right). Our ANOVA test showed a significant difference in the set with $F(2, 750) = 9.48, p < 0.001$). The post-hoc Tukey HSD tests revealed a significant statistical difference between N1 ($M = 76.36, SD = 16.51$) and N2 ($M = 70.66, SD = 16.30$) as well as N1 and N7 ($M = 70.77, SD = 16.87$). N2 and N7 showed no significant difference between each other. The difference in SUS score puts the default password generator (N2) and the novel gamified avatar nudge (N7) in the adjective rating bracket "OK", not quite reaching the dynamic password meter's (N1) "Good" rating. All nudges fulfill the highest acceptability rating of "Acceptable", as rated by Brooke et al. [4, p. 20].

5.2.4 Effect of Gaming Frequency. In our work we are especially interested in the effect of gaming frequency on the acceptance and effectiveness of our nudges. Regarding the SUS scores in general, we found no significant differences when factoring in the levels of gaming frequency ($F(4, 748) = 1.90, p = 0.11$). For password strength, however, we did detect a significant effect on the capped zxcvbn-score from 1 to 4 ($F(4, 995) = 2.72, p = .029$). When looking at the post-hoc Tukey HSD test, we can see that the significant difference appears between "weekly" and "never", where the zxcvbn-score for "weekly" is $M = 3.08, SD = 1.08$ and for "never" it is $M = 1.84, SD = 1.04$. When comparing the strength regarding log(*guesses*), there is no significant difference between the gaming groups ($F(4, 253) = 0.94, p = .441$).

While the general SUS scores and strength measurements give insights about the general assessment of all nudges, we intend to evaluate differences in acceptance and effectiveness of gamified nudges versus non-gamified nudges, comparing groups with different gaming habits. Thus, we looked at how specific nudges performed for those groups. Regarding the novel gamified avatar nudge (N7), all participants assigned to that nudge rated its usability similarly, independently of their gaming frequency ($F(4, 253) = 0.51, p = .732$). For all other nudges, we did not detect significant differences either. We performed the same analysis concerning password strength. For none of the test conditions (N1, N2, N7, control group) the password strength differed between groups of gaming frequencies (e.g. for N2: $F(4, 251) = 0.16, p = .960$). Hence, for nudge usability or effectiveness it made no difference if participants with a specific gaming frequency were assigned to a gamified or non-gamified nudge.

5.2.5 Comparison of Short-term Memorability. We tracked the amount of failed tries to recall the password, limited at 4. An amount of 0 failed tries declares that the user recalled their password immediately. We performed ANOVA tests over all test conditions, including the subgroups N2.d and N2.c and found a significant difference: $F(4, 995) = 74.59, p < 0.001$. A post-hoc Tukey HSD test showed that only N2.d differs significantly from all other groups, with its memorability being much worse than that of all other test groups where users created their own passwords (e.g., comparing N2.d with N1 the difference is highly significant with $p < .001$ and a mean difference of -0.63). To sum it up, all evaluated nudges except for the default password generator when using the default option performed similarly regarding memorability. None of the other nudges explicitly facilitated or hindered memorability in comparison to the control group without a nudge.

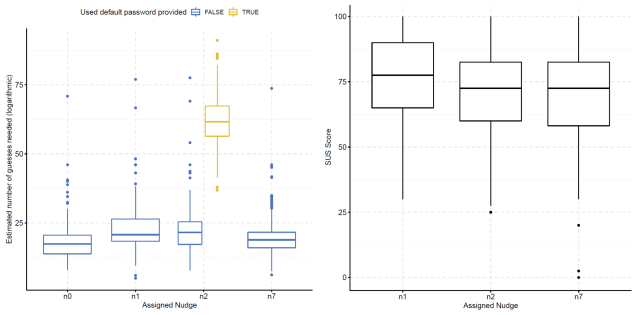

Figure 2: Left: strength of passwords per condition in estimated number of guesses needed (logarithmic). Right: SUS score per nudge.

6 DISCUSSION OF RESULTS AND LIMITATIONS

The goal of this work is to evaluate the potential of gamified nudges in comparison to non-gamified nudges and state-of-the art concepts to contribute to the pool of effective nudges and facilitate personalization in cybersecurity. We have identified a lack of research regarding large-scaled evaluations of concrete nudge implementations, comparing novel nudges against state-of-the-art nudges such as the dynamic password meter of Ur et al. [50], as well as regarding gamification in nudges for cybersecurity. To make initial efforts towards filling this gap, we iteratively designed novel gamified nudges for password strength, in particular an interactive, customizable avatar nudge guiding the user through the registration process using natural language cues, and compared their usability, effectiveness and short-term memorability to other (established) nudges without gamification elements. Our evaluation took place in a two-fold study design, consisting of a preliminary think-aloud study for qualitative insights, followed up by a large-scaled online experiment with 1,000 participants to quantitatively compare usability, effectiveness, and memorability. Notably, we asked about the gaming frequency of our participants, as we hypothesized that gaming frequency affects the acceptance of our gamified nudge. We gained some interesting insights, contributing to the human-centered research in password nudges:

We found that *(1)* while most evaluated nudges were having rather small but significant effects on password strength, our gamified avatar nudge N7 was less effective than the other nudges. Contrary to our expectations, that finding holds true *(2)* also when considering only participants that frequently play online games and, thus, may feel more appealed to gamification. While Jeske et al. [22] among others suggest that user differences play a role in security decision making, we could not confirm that for gaming preferences in our setting. Among the evaluated sample, the generated default passwords under condition N2.d achieved the best results regarding password strength, followed by the dynamic password meter in the style of Ur et al. [50]. Interestingly *(3)*, the default generator nudge was comparably effective regarding password strength also when participants chose to not use the suggested password but created their own under condition N2.c.

Memorability is a crucial factor for password nudges as well. Due to the limitations of our online experiment, we decided to evaluate short-term memorability to gain first comparable insights. While the default generated passwords in condition N2.d yielded great results concerning password strength, the nudge comes with an essential shortcoming in short-term memorability. Indeed *(4)*, it performed significantly worse than all other evaluated nudges, suggesting that our short-term memorability study can not reproduce the high memorability results of Ghazvininejad and Knight [17]. This finding is not surprising, as the participants did not have to interact with the password at all while executing the registration process. In all other test conditions, the participants were forced to invest a minimum of thoughts into the password creation which yielded better short-term memorability. While Furnell et al. [14] suggest gamification as an enhancing factor for memorability, our avatar nudge did not appear to be effective concerning that matter.

Our study revealed interesting insights into how participants assessed different nudges with and without gamification elements. During our large-scale online experiment, we found *(5)* that the state-of-the-art dynamic password meter nudge N1 performed significantly better than our gamified nudge N7 and the default password generator nudge N2 while N2 and N7 did not significantly differ from each other regarding usability. We assume this to be caused by N1 being a well-established, state-of-the-art nudge, as multiple participants of the preliminary study praised it for being a *"known environment"* (e.g., P #970, P #140, P #493). However *(6)*, all evaluated nudges achieved satisfying SUS scores.

To sum up our findings, we propose the following answers to our research questions:

(Q1): How do gamified nudges compare to a state-of-the-art password meter?

The gamified avatar as a novel nudge did not encourage password strength comparably to the dynamic password meter by Ur et al.[50]. Still, the avatar nudge can be considered an effective nudge as it indeed enhanced password strength compared to the control group without a nudge. Regarding usability, the avatar nudge yielded sufficient results, however, it was outperformed by the usability of the state-of-the-art dynamic password meter.

(Q2): Does a user's gaming frequency increase the acceptance of a gamified nudge?

Contrarily to our expectations and the results of other studies (e.g., [20]), the users' gaming frequency had no effect on neither the acceptance nor the effectiveness of gamified versus non-gamified nudges.

Limitations and Future Work: While the study we report is a first step towards extending the pool of effective nudges in cybersecurity for personalization, it has bears limitations. The evaluation of a nudge in an online experiment does not lead to information about efficacy as realistic as in a real-world scenario. Future studies may complement evaluations on the efficacy and memorability of the gamified nudge by using a long-term study design, also including more participants of lower education. While our work is a first step to assess the potential of gamified versus non-gamified nudges for cybersecurity on a short-term basis, we cannot make a final conclusion about their overall efficacy before testing them under more realistic circumstances. Hence, we suggest to utilize other techniques of data collection in the future. Additionally, we propose

that individualizing nudges with regard to the intended audience of a service (e.g. readers of an information security news feed possibly being more tech-savvy) could be an easier and less costly implementation of individualizing nudges than identifying individual user traits by questionnaire. This could lead to greater acceptance and efficacy of the shown nudges, as discussed by Knijnenburg [28].

7 CONCLUSION

In this work, we examined a sample of nudges for stronger passwords, focusing on a novel gamified nudge in comparison to state-of-the-art concepts and other nudges without gamification elements. We evaluated the nudges in a two-fold study design, conducting a qualitative think-aloud preliminary study and a large-scaled online experiment with 1,000 participants. Considering password strength, a usability score, and short-term memorabilty, we found that users still tend to prefer well-known concepts for password creation assistance. While our gamified nudge showed significant effects on password strength compared to a control group without a nudge, the effects are rather small. Interestingly, we further found that the gaming frequency of our participants made no difference on the effect of our gamified nudge.

ACKNOWLEDGMENTS

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1119 (CROSSING) – 236615297.

REFERENCES

[1] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, et al. 2017. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys (CSUR)* 50, 3 (2017), 1–41.

[2] Mahdi Nasrullah Al-Ameen, Matthew Wright, and Shannon Scielzo. 2015. Towards making random passwords memorable: Leveraging users' cognitive ability through multiple cues. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2315–2324.

[3] Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do Multilingual Users Prefer Chat-Bots That Code-Mix? Let's Nudge and Find Out! *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 041 (May 2020), 23 pages. https://doi.org/10.1145/3392846

[4] John Brooke. 1996. SUS - A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7. www.TBIStaffTraining.info

[5] Pete Burnap, Robert Carolina, Awais Rashid, M. Angela Sasse, Carmela Troncoso, Wenke Lee, Gianluca Stringhini, Herve Debar, Vassil Roussev, Nigel Smart, and et al. 2019. *The Cyber Security Body of Knowledge* (1 ed.).

[6] Gamze Canova, Melanie Volkamer, Clemens Bergmann, and Roland Borza. 2014. NoPhish: An anti-phishing education app. *Lecture Notes in Computer Science* 8743 (2014), 88–192. https://doi.org/10.1007/978-3-319-11851-2_14

[7] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. (2019), 1–15. https://doi.org/10.1145/3290605.3300733

[8] Lucian Constantin. 2011. Sony Stresses that PSN Passwords Were Hashed. http://news.softpedia.com/news/SonyStresses-PSN-Passwords-Were-Hashed-198218.shtml. http://news.softpedia.com/news/SonyStresses-PSN-Passwords-Were-Hashed-198218.shtml Accessed: 2021-01-18.

[9] Matteo Dell'Amico, Pietro Michiardi, and Yves Roudier. 2010. Password strength: An empirical analysis. In *2010 Proceedings IEEE INFOCOM*. IEEE, 1–9.

[10] Janna Lynn Dupree, Richard Devries, Daniel M. Berry, and Edward Lank. 2016. Privacy personas: clustering users via attitudes and behaviors toward security practices. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM* (2016), 5228–5239. https://doi.org/10.1145/2858036.2858214

[11] Serge Egelman and Eyal Peer. 2015. The myth of the average user: Improving privacy and security systems through individualization. *Proceedings of the 2015 New Security Paradigms Workshop* (2015), 16–28. https://doi.org/10.1145/2841113.2841115 arXiv:arXiv:1508.06655v1

[12] Sascha Fahl, Marian Harbach, Yasemin Acar, and Matthew Smith. 2013. On the Ecological Validity of a Password Study. In *Proceedings of the Ninth Symposium on Usable Privacy and Security* (Newcastle, United Kingdom) *(SOUPS '13)*. Association for Computing Machinery, New York, NY, USA, Article 13, 13 pages. https://doi.org/10.1145/2501604.2501617

[13] Marsha E. Fonteyn, Benjamin Kuipers, and Susan J. Grobe. 1993. A Description of Think Aloud Method and Protocol Analysis. *Qualitative Health Research* 3, 4 (Nov. 1993), 430–441. https://doi.org/10.1177/104973239300300403 Publisher: SAGE Publications Inc.

[14] Steven Furnell, Faisal Alotaibi, and Rawan Esmael. 2019. Aligning Security Practice with Policy : Guiding and Nudging towards Better Behavior. *Proceedings of the 52nd Hawaii International Conference on System Sciences* 6 (2019), 5618–5627.

[15] Steven M. Furnell, Adila Jusoh, and Dimitris Katsabas. 2006. The challenges of understanding and using security: A survey of end-users. *Computers & Security* 25, 1 (2006), 27–35.

[16] Simson Garfinkel and Heather Richter Lipford. 2014. Usable security: History, themes, and challenges. *Synthesis Lectures on Information Security, Privacy, and Trust* 5, 2 (2014), 1–124.

[17] Marjan Ghazvininejad and Kevin Knight. 2015. How to Memorize a Random 60-Bit String. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, 1569–1575. https://doi.org/10.3115/v1/N15-1180

[18] Paul A. Grassi, James L. Fenton, Elaine M. Newton, Ray A. Perlner, Andrew R. Regenscheid, William E. Burr, Justin P. Richer, Naomi B. Lefkovitz, Jamie M. Danker, Yee-Yin Choong, Kristen K. Greene Theofanos, and Mary F. 2017. *NIST Special Publication 800-63B, Digital Identity Guidelines*. Technical Report. NIST. https://pages.nist.gov/800-63-3/sp800-63b.html

[19] Marian Harbach, Markus Hettig, Susanne Weber, and Matthew Smith. 2014. Using Personal Examples to Improve Risk Communication for Security & Privacy Decisions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 2647–2656. https://doi.org/10.1145/2556288.2556978

[20] Katrin Hartwig and Christian Reuter. 2020. Dealing with Transparency in Nudges: Nudging Users Towards Stronger Passwords by Using Transparent Visualizations. *Behaviour & Information Technology* (2020).

[21] Katrin Hartwig and Christian Reuter. 2021. Nudge or Restraint: How do People Assess Nudging in Cybersecurity - A Representative Study in Germany. In *The 2021 European Symposium on Usable Security (EuroUSEC)*.

[22] Debora Jeske, Lynne Coventry, and Pam Briggs. 2014. Nudging whom how : IT proficiency , impulse control and secure behaviour. *Proceedings of the CHI Workshop on Personalizing Behavior Change Technologies* April (2014), 1–4.

[23] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

[24] Jeffrey P. Kaleta, Jong Seok Lee, and Sungjin Yoo. 2019. Nudging with construal level theory to improve online password use and intended password choice. *Information Technology & People* (2019).

[25] Shipi Kankane, Carlina Dirusso, and Christen Buckley. 2018. Can we nudge users toward better password management ? An initial study. *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing System. ACM* (2018), 1–6.

[26] Kianoosh Karimi and Javad Nickpayam. 2017. Gamification from the Viewpoint of Motivational Theory. *Emerging Science Journal* 1, 1 (2017), 34–42. https://doi.org/10.28991/esj-2017-01114

[27] Johannes Kiesel, Benno Stein, and Stefan Lucks. 2017. A Large-scale Analysis of the Mnemonic Password Advice.. In *NDSS*.

[28] Bart Knijnenburg. 2017. Privacy? I Can't Even! Making a Case for User-Tailored Privacy. *IEEE Security and Privacy* 15, 4 (2017), 62–67. https://doi.org/10.1109/MSP.2017.3151331

[29] Jonna Koivisto and Juho Hamari. 2019. The rise of motivational information systems: A review of gamification research. *International Journal of Information Management* 45 (2019), 191–210.

[30] Saranga Komanduri, Richard Shay, Patrick Gage Kelley, Michelle L Mazurek, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Serge Egelman. 2011. Of passwords and people: measuring the effect of password-composition policies. In *Proceedings of the sigchi conference on human factors in computing systems*. 2595–2604.

[31] Dimosthenis Kotsopoulos, Cleopatra Bardaki, Thanasis G Papaioannou, Katerina Pramatari, and George D Stamoulis. 2020. User-Centered Gamification. *International Journal of E-Services and Mobile Applications* 12, 2 (2020), 15–39. https://doi.org/10.4018/ijesma.2020040102

[32] Nicholas Micallef and Nalin Asanka Gamagedara Arachchilage. 2017. A Serious Game Design : Nudging Users ' Memorability of Security Questions. In *Australasian Conference on Information Systems*. 1–11.

[33] Randall Munroe. 2011. *XKCD #936: "Password Strength"*. https://xkcd.com/936/ Accessed: 2021-01-18.
[34] Michael I. Norton, Daniel Mochon, and Dan Ariely. 2012. The IKEA effect: When labor leads to love. *Journal of Consumer Psychology* 22, 3 (July 2012), 453–460. https://doi.org/10.1016/j.jcps.2011.08.002
[35] Jacques Ophoff and Frauke Dietz. 2019. Using Gamification to Improve Information Security Behavior: A Password Strength Experiment. *IFIP Advances in Information and Communication Technology* 557 (2019), 157–169. https://doi.org/10.1007/978-3-030-23451-5_12
[36] Sarah Pearman, Shikun Aerin Zhang, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2019. Why people (don't) use password managers effectively. *Fifteenth Symposium on Usable Privacy and Security (SOUPS)* (2019). https://www.usenix.org/conference/soups2019/presentation/pearman
[37] Eyal Pe'er, Serge Egelman, Marian Harbach, Nathan Malkin, Arunesh Mathur, and Alisa Frik. 2019. Nudge Me Right: Personalizing Online Nudges to People's Decision-Making Styles. *SSRN Electronic Journal* (2019). https://doi.org/10.2139/ssrn.3324907
[38] Kimberly Perzel and David Kane. 1999. Usability patterns for applications on the world wide web. In *Proceedings of the Pattern Languages of Programming Conference*, Vol. 99.
[39] Karen Renaud, Verena Zimmerman, Joseph Maguire, and Steve Draper. 2017. Lessons Learned from Evaluating Eight Password Nudges in the Wild. In *The LASER Workshop: Learning from Authoritative Security Experiment Results (LASER 2017)*. USENIX Association, 25–37. https://www.usenix.org/conference/laser2017/presentation/renaud
[40] Karen Renaud and Verena Zimmermann. 2018. Nudging folks towards stronger password choices: providing certainty is the key. *Behavioural Public Policy* 3, 02 (2018), 228–258. https://doi.org/10.1017/bpp.2018.3
[41] Karen Renaud, Verena Zimmermann, Joseph Maguire, and Steve Draper. 2017. Lessons Learned from Evaluating Eight Password Nudges in the Wild.
[42] Eva Ruegenhagen and Bernard Rummel. 2013. System Usability Scale–jetzt auch auf Deutsch. *SAP Global Design Enablement Team* (2013).
[43] Sam Scholefield and Lynsay A Shepherd. 2019. Gamification Techniques for Raising Cyber Security Awareness. In *International Conference on HCI for Cybersecurity, Privacy and Trust (HCI-CPT)*.
[44] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. 2007. Anti-Phishing Phil: The design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd Symposium on Usable Privacy and Security (SOUPS '07)*, Vol. 229. 88–99. https://doi.org/10.1145/1280680.1280692
[45] Mario Silic and Paul Benjamin Lowry. 2019. Using Design-Science Based Gamification to Improve Organizational Security Training and Compliance. *Journal of Management Information Systems (JMIS)* (2019).
[46] Tetsuji Takada and Yumeji Hattori. 2020. Giving Motivation for Using Secure Credentials through User Authentication by Game. In *Proceedings of the International Conference on Advanced Visual Interfaces*. ACM, New York, NY, USA, 1–3. https://doi.org/10.1145/3399715.3399950
[47] Richard H. Thaler and Cass R. Sunstein. 2009. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.
[48] Iis Tussyadiah, Shujun Li, and Graham Miller. 2019. Privacy protection in tourism: Where we are and where we should be heading for. *Information and Communication Technologies in Tourism* (2019), 278–290. https://doi.org/10.1007/978-3-030-05940-8_22
[49] Blase Ur, Patrick Gage Kelley, Saranga Komanduri, Joel Lee, Michael Maass, Michelle L. Mazurek, Timothy Passaro, Richard Shay, Timothy Vidas, Lujo Bauer, et al. 2012. How does your password measure up? the effect of strength meters on password creation. In *Presented as part of the 21st {USENIX} Security Symposium ({USENIX} Security 12)*. 65–80.
[50] Blase Ur, Gloria Mark, Susan Fussell, Cliff Lampe, Juan Pablo Hourcade, Caroline Appert, Daniel Wigdor, Felicia Alfieri, Maung Aung, Lujo Bauer, Nicolas Christin, Jessica Colnago, Lorrie Faith Cranor, Henry Dixon, Pardis Emami Naeini, Hana Habib, Noah Johnson, and William Melicher. 2017. Design and Evaluation of a Data-Driven Password Meter. (2017), 3775–3786. https://doi.org/10.1145/3025453.3026050
[51] Thomas Waltemate, Dominik Gall, Daniel Roth, Mario Botsch, and Marc Erich Latoschik. 2018. The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. *IEEE transactions on visualization and computer graphics* 24, 4 (2018), 1643–1652.
[52] Daniel Lowe Wheeler. 2016. zxcvbn: Low-budget password strength estimation. In *25th USENIX Security Symposium (USENIX Security 16)*. 157–173.
[53] Luke Wroblewski. 2008. *Web form design: filling in the blanks*. Rosenfeld Media.
[54] Verena Zimmermann and Nina Gerber. 2020. The password is dead, long live the password–A laboratory study on user perceptions of authentication schemes. *International Journal of Human-Computer Studies* 133 (2020), 26–44.

A APPENDIX

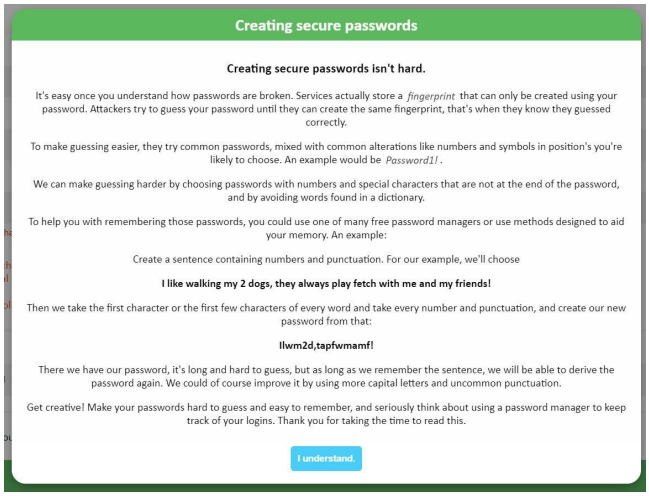

Figure 3: Screenshot of help text (English version. German variant used in online study)

Table 2: Preliminary Study: Participants were asked to rate how much they agree with a given statement on a range of 1 to 5, with 1 representing "Strongly disagree" and 5 representing "Strongly agree". The questions differed between nudges.

Nudge-specific questions
Nudge 1 - Dynamic Password Meter "I found it clear which requirements were fulfilled" "I found it easy to distinguish between mandatory and optional requirements" "The visualization helped me improve on my password"
Nudge 2 - Default password "I find the generated password easy to remember" "I would use similarly generated passwords in real-life services."
Nudge 3 - Radar chart "I found the chart easy to understand" "I found it easy to improve my password with the feedback provided"
Nudge 4 - Expectation and Reflection "I found it easy to approximate the strength of my password" "Having to think about the strength of my password prompted me to improve on it"
Nudge 5 - Reminding of the Consequences "Being reminded of the consequences helped me reflect on my password strength" "I improved my password strength after being reminded of the consequences"
Nudge 6 - Chat "I found the chat messages easy to understand" "It was clear to me what input was required at any time" "The provided feedback helped me improve on my password"
Nudge 7 - Avatar "The creation of an avatar was easy and intuitive" "I valued the feedback given by the avatar" "The feedback provided was clear and easily understandable" "The feedback provided helped me improve on my password" "I would like to see this visualization adopted by more services" "I had enough options for customization"

Nudge	**Strength**		**N**
	M	*SD*	
Control Group	18.05	47.91	247
N1 - Dynamic Meter	22.93	69.55	239
N2 - Default Password	53.79	341.03	256
N2.d - Default Password *(used default)*	62.02	77.07	200
N2.c - Default Password *(used custom)*	24.40	176.27	56
N7 - Avatar	20.08	51.73	258

Table 3: Password strength in log(guesses)