Behaviour & Information Technology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tbit20

Nudging users towards better security decisions in password creation using whitebox-based multidimensional visualisations

Katrin Hartwig & Christian Reuter

To cite this article: Katrin Hartwig & Christian Reuter (2021): Nudging users towards better security decisions in password creation using whitebox-based multidimensional visualisations, Behaviour & Information Technology, DOI: 10.1080/0144929X.2021.1876167

To link to this article: https://doi.org/10.1080/0144929X.2021.1876167

Published online: 25 Jan 2021.

BEHAVIOUR & INFORMATION TECHNOLOGY
https://doi.org/10.1080/0144929X.2021.1876167

Taylor & Francis
Taylor & Francis Group

Nudging users towards better security decisions in password creation using whitebox-based multidimensional visualisations

Katrin Hartwig and Christian Reuter

Science and Technology for Peace and Security (PEASEC), Technische Universität Darmstadt, Darmstadt, Germany

ABSTRACT

Nudging users to keep them secure online has become a growing research field in cybersecurity. While existing approaches are mainly blackbox based, showing aggregated visualisations as one-size-fits-all nudges, personalisation turned out promising to enhance the efficacy of nudges within the high variance of users and contexts. This article presents a disaggregated whitebox-based visualisation of critical information as a novel nudge. By segmenting users according to their decision-making and information processing styles, we investigate if the novel nudge is more effective for specific users than a common black-box nudge. Based on existing literature about critical factors in password security, we designed a dynamic radar chart and parallel coordinates as disaggregated visualisations. We evaluated the short-term effectiveness and users' perception of the nudges in a think-aloud prestudy and a representative online evaluation (N=1.012). Our findings suggest that dynamic radar charts present a moderately effective nudge towards stronger passwords regarding short-term efficacy and are appreciated particularly by players of role-playing games.

ARTICLE HISTORY
Received 7 April 2020
Accepted 5 January 2021

KEYWORDS
Nudging; whitebox; usable security; personalisation; passwords

1. Introduction

When forcing people 's decisions towards a desired outcome without being convenient, people tend to find workarounds. For example, if users are forced to adopt higher online security, it may reduce their willingness to follow the advice when the benefits are not clear and the desired behaviour appears to be a disproportionately big effort (Merkel and Wiczorek 2012; Jeske, Coventry, and Briggs 2014a). In that case, users often choose convenience over security. Instead of forcing, nudging is a promising concept to steer people's behaviour. A nudge is '*any aspect of the choice architecture that alters people's behaviour in a predictable way without forbidding any options or significantly changing their economic incentives*' (Thaler and Sunstein 2009, p. 6). However, Renaud and Zimmerman (2017) include nudges as well that come with a maximisation of non-monetary utility, introducing the term of 'enriched nudges'. Nudges can be applied in different ways. For instance, the advertisement industry uses nudges to convince people of buying a specific product. More favourably, applications in healthcare try to nudge people towards a healthier life. As it is a very subjective perception of what is best for people, ethical guidelines have to be considered. Renaud and Zimmermann (2018a) have derived general principles for ethical nudging from the existing literature. For instance, they state that nudges should be transparent to the nudgees and should only be used when the benefits are clear. Other researchers argue that nudges do not negatively affect autonomy and that transparency does also not increase it (Wachner, Adriaanse, and De Ridder 2020).

Nudging is a long-established concept used in many contexts, such as the health sector. Interestingly, research has more recently started to consider nudging also for application areas such as the assessment of news credibility in social media (Bhuiyan et al. 2018) or to facilitate social interactions (Chen and Abouzied 2016). Meanwhile, it has become a research field in cybersecurity and privacy as well (Acquisti et al. 2017). As the behaviour of human end-users is considered central within the cybersecurity chain (Herbert, Schmidbauer-Wolf, and Reuter 2020; Biselli and Reuter 2021), design in terms of 'usable security' has become more and more crucial. For instance, researchers have evaluated the effects of security awareness delivery methods (Abawajy 2012). Others aim to bridge the gap between security and usability through mental models (Mohamed, Chakraborty, and Dehlinger 2017). Recent works have started to investigate how bad security and privacy decisions of users can be nudged towards more beneficial decision-making

CONTACT Katrin Hartwig hartwig@peasec.tu-darmstadt.de Science and Technology for Peace and Security (PEASEC), Technische Universität Darmstadt, Pankratiusstraße 2, Darmstadt 64289, Germany

(Acquisti et al. 2017). A very common context is the exploration of measurements against phishing, considering the influence of user characteristics and behaviour tendencies (e.g. Jansson and Von Solms 2013; Li et al. 2014; Ramesh, Selvakumar, and Venugopal 2017; Kim, Lee, and Kim 2019; House and Raja 2019). As password creation is a critical process as well, where users still often fail to create safe passwords, they need to be provided with more security guidance (Huh et al. 2017). For instance, Segreti et al. (2017) investigated if adaptive password-composition policies can nudge users to create usable and secure passwords in a large-scaled online experiment. Currently, there are mainly one-size-fits-all nudges in cybersecurity although there is a high variance of users and situations (Peer et al. 2019). To address that flaw, personalisation is considered promising by current research (Savola and Heinonen 2011; Knijnenburg 2017; Peer et al. 2019). For example, Peer et al. (2019) point out that personalised nudges have stronger outcomes and reduce the risk of individuals reacting contrarily to the desired behaviour. Therefore, it is necessary to provide distinct user groups with a targeted type of nudge. To achieve that, on the one hand, user groups need to be segmented reasonably. On the other hand, it is necessary to investigate which type of nudge is most effective for each identified user group.

Often, cybersecurity measures are hard to understand for the average user. Interface design can be used to address that problem, for example by providing visual feedback on decisions in critical situations or displaying relevant information (Boyce et al. 2011). When visualising information, people 's preferences regarding transparency differ. While for some people it is sufficient to see an aggregated output of a calculation, others will not trust the output unless they can comprehend how it was generated. For the latter, algorithmic transparency is essential. There are mainly two different approaches to explain algorithms: black-box and white-box approaches (Cheng et al. 2019). In a black-box approach, the user can observe the input and the output but not what happens in between. As a result, that may lead to reactance for those who need more information. On the other hand, black-box approaches are less likely to cause information overload. White-box approaches enable the user to understand how the output was generated. Hence, the underlying basis for output generation is transparent to the user. That is beneficial especially for individuals that are otherwise likely to feel reactance or mistrust (Hartwig and Reuter 2019, 2020). However, others might feel overwhelmed by too much information (Kaufhold et al. 2020). Explainable algorithms have emerged to be a highly relevant field of research, for instance, to understand how decisions in health and finance are made by machine learning techniques (e.g. Abdul et al. 2018; Cheng et al. 2019; Wang et al. 2019). In the context of nudging, calculations are commonly based on less complex methods. Therefore, giving transparent (and, thus, whitebox-based) reasons can often be accomplished by simply disaggregating the calculated information and accordingly visualising relevant dimensions in a comprehensible way.

For now, most nudges are based on a unidimensional visualisation of aggregated information. For instance, password meters usually show aggregated results of metrics using colour code. Those nudges only tell if a password is weak but not why. Hence, they are black-box-based and do not let users know what to do to improve their passwords. However, gaining a comprehensive picture by showing a more comprehensive collection of information instead of aggregated results can help users to make better decisions (Savola and Heinonen 2011; Ur et al. 2016). White-box approaches can facilitate an understanding of why the output was generated by showing multiple dimensions in one visualisation. Thus it is not only transparent to the users on what dimensions the calculation of password strength is based but also what to do to improve their password. Therefore, it is a critical field of research to investigate if multidimensional visualisations are more effective nudges for specific user groups than aggregated visualisations and can be considered a central personalised component.

The scientific contribution of this article is the evaluation of two different white-box nudges in comparison to a simple black-box password meter in the context of password creation. The white-box nudges make transparent how password strength was calculated and what the user can do to improve the password. The black-box meter on the other hand does not make transparent to the user how password strength was calculated and, thus, does not provide information on what to do to improve password strength. For personalisation, we aim to investigate if specific user groups prefer white-box-based visual feedback over current unidimensional and mostly blackbox-based visualisations. To evaluate the novel nudge, we focus on two exemplary visualisation techniques: radar charts and parallel coordinates. Radar charts are commonly used in role-playing games to visualise a characters' strengths and weaknesses, as shown in Figure 1, and therefore might be a familiar sight with motivating influences for several user groups.

Hence, our research questions are:

How can whitebox-based multidimensional visualisations provide an effective nudge towards better security

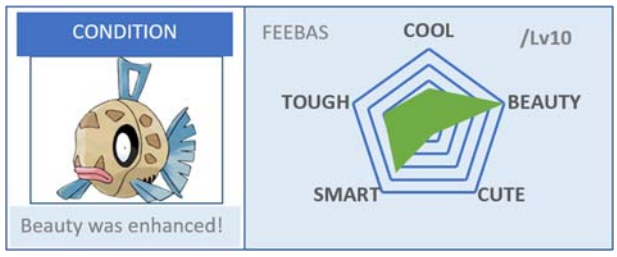

Figure 1. Radar chart visualising the strengths of a Pokemon (https://bulbapedia.bulbagarden.net/wiki/Feebas).

decisions in password creation for specific user groups? Do users feel overwhelmed by a more comprehensive and multidimensional visualisation of information in the context of security decisions in password creation?

To answer our research questions, we iteratively conducted a prestudy with individual sessions using the think-aloud method and a representative evaluation (N=1.012) including a survey and an online experiment. We applied user-centred design methods as they have proven to be efficient and effective for cybersecurity visualisation design (Boyce et al. 2011; McKenna, Staheli, and Meyer 2015). As a nudge, we implemented a dynamic visualisation of password strength during password creation. To contribute to the idea of personalised nudges, we segmented our user groups by applying two standardised psychometric tests. The article is organised as follows: Section 2 presents related work on nudging in cybersecurity. We will conclude Section 2 by highlighting a research gap. In Section 3, we present an overview of our research design of a novel potential nudge in cybersecurity. In Section 4, our prestudy and its implications for further improvements of the nudge are introduced. In Section 5, we present a representative evaluation, proposing the concept and implementation of a whitebox-based multidimensional visualisation as a novel nudge in cybersecurity, as well as the evaluation method and its results. We discuss the contributions of our approach in Section 6. Finally, we point out limitations and potential for future work in Section 7.

2. Related work and research gap

Within the research area of nudging, we can identify several current trends. Particularly, personalisation instead of one-size-fits-all nudges and nudging through visualisation of information have emerged as promising trends and are addressed by current research. In the following, we will give an overview of concrete studies on nudging, focusing on the context of cybersecurity. To get an idea about commonly used nudges in cybersecurity, we present an excerpt of existing nudges or nudging attempts for stronger passwords which have been evaluated in current research in Figure 2(a): indicating by the length of a dachshund how long a password will be valid ('enriched nudge') (Renaud and Zimmerman 2017), (b): an image of eyes that aims to remind social norms for stronger passwords but failed as a nudge (Renaud et al. 2017), (c): a password meter and a list with suggested improvements (Ur et al. 2017). While the password has often been criticised as authentication scheme, it is still preferred by many users (Zimmermann and Gerber 2020). In a laboratory study, Zimmermann and Gerber (2020) compared participants' ratings of different authentication schemes and found that the password scored highest regarding preference, usability and intention to use. However, the authors recommend providing better guidance in secure password creation, with nudging being a viable option.

2.1. Personalised nudges

Studies have pointed out that nudges in the context of cybersecurity are often not as effective as desired. For example, Kankane, Dirusso, and Buckley (2018) examined the effectiveness of different nudges for password management. They found that none of the examined nudges was effective enough to significantly change the individuals' behaviour concerning password creation. According to Egelman and Peer (2015), many security and privacy systems are made usable only for the average user. As a result, compliance is limited to certain end-users while others do not benefit. However, when designed for an individual, compliance is likely to improve (Egelman and Peer 2015). To address that problem, a recent trend in nudging is personalisation. Several researchers suggest showing personalised nudges according to user traits instead of one-size-fits-all nudges (Jeske, Coventry, and Briggs 2014b; Egelman and Peer 2015; Knijnenburg 2017; Renaud et al. 2017; Peer et al. 2019; Tussyadiah, Li, and Miller 2019). For instance, Knijnenburg (2017) argue that showing tailored nudges can support users in making better privacy decisions.

To provide individuals with the subjectively most effective nudge, distinct user groups need to be identified. There are various approaches to segment users. For example, Dupree et al. (2016, p. 5228) identified five clusters: *'Fundamentalists, Lazy Experts, Technicians, Amateurs and the Marginally Concerned'*. Further, Egelman and Peer (2015) suggest segmenting users according to their decision-making styles and risk-taking attitudes to predict privacy and security behaviour and accordingly show the best fitting nudge. More recently, Egelman, Harbach, and Peer (2016) followed up on that idea by performing online

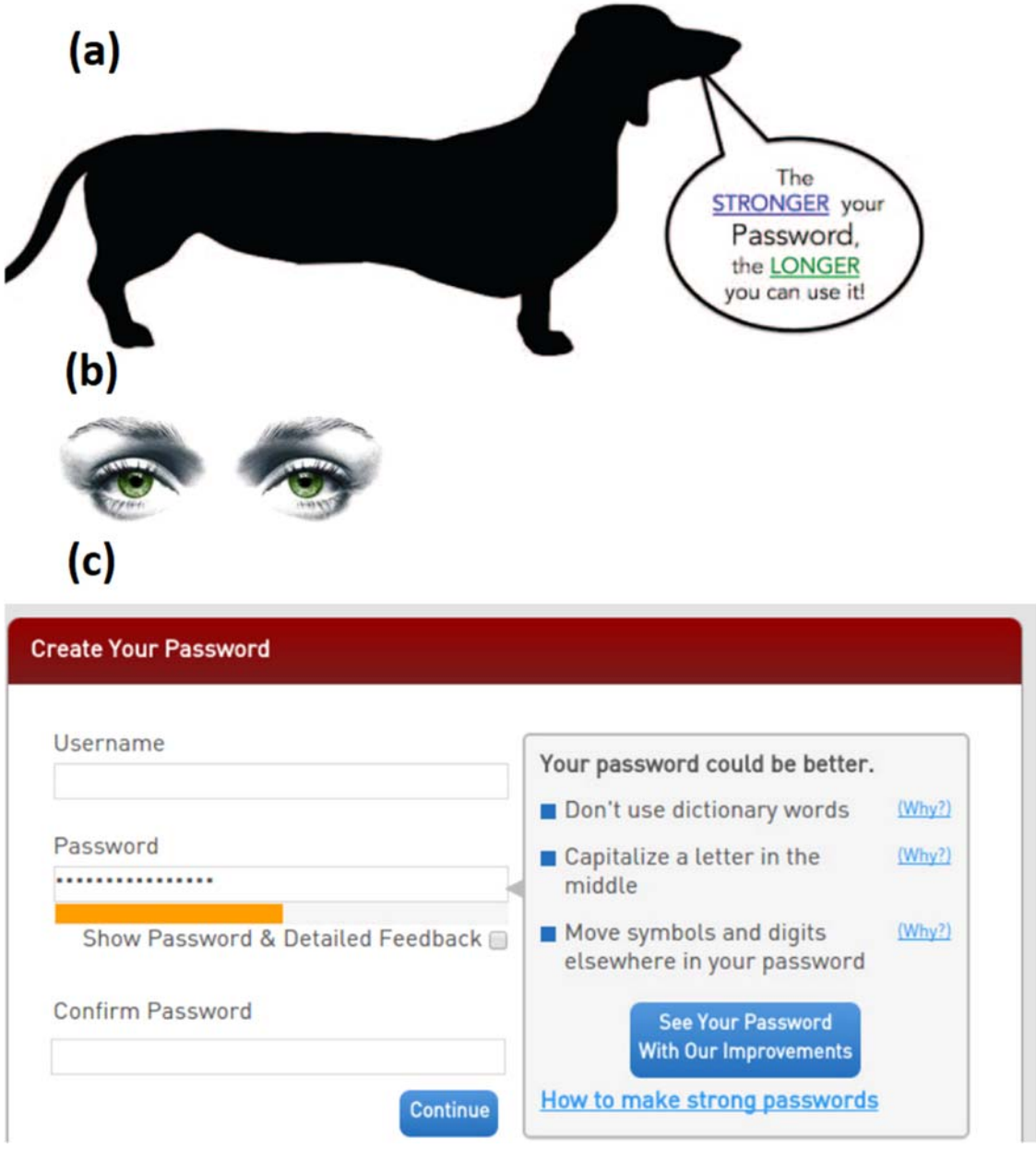

Figure 2. Several existing nudges and nudging attempts for stronger passwords (a) Renaud and Zimmerman (2017), (b) Renaud et al. (2017), (c) Ur et al. (2017).

experiments. The results suggest that the Security Behaviour Intentions Scale (SeBIS) does indeed predict certain computer security behaviours. Personalisation is also an upcoming challenge in the context of browser warnings, as Reeder et al. (2018) argue.

While personalisation of nudges is considered promising by several studies, only a few have implemented the concept within the cybersecurity context. Peer et al. (2019) tested people's decision-making styles to personalise nudges for stronger passwords in two online experiments (N=2047). They argue that choosing a nudge from a pool of multiple existing nudges could be more effective than showing the same nudge to everyone. Choosing from five frequently used nudges (e.g. feedback on how long it takes to crack the password) according to the decision-making style, Peer et al. found that decision-making styles can indeed be used to personalise nudges. They achieved stronger passwords with personalisation than with one-size-fits-all nudges (Peer et al. 2019). Furthermore, Jeske, Coventry, and Briggs (2014b) examined if the effectiveness of nudges depends on user characteristics such as impulse control when selecting a public wireless network, asking 104 students in a role-play to select a network given a specific nudge. They found that user differences indeed play a role in security decision-making. Hence, those

results suggest personalisation of nudges in cybersecurity is a promising trend which should be striven for in future research (Jeske, Coventry, and Briggs 2014b). To allow many individuals to benefit, it is necessary to create a pool of different nudges.

2.2. Nudging through transparent v isualisation

Besides personalisation, another trend for nudging in cybersecurity is visualising information in a comprehensible way. Visualisation can be considered a very valuable way to provide information as it provides *'the highest bandwidth channel from computer to the human'* (Ware 2012, p. 2). Boyce et al. (2011), among others, argue that it is critical to make cybersecurity measures easier to understand, for instance using features of the user interface.

Current studies on nudging in cybersecurity focus mostly on uni- or two-dimensional visualisations, for example by displaying black-box password meters. Ur et al. (2016, p. 3757) state that *'targeted, data-driven feedback during password creation'* is promising to assist users in creating stronger passwords. Hence, password meters appear to be an appropriate concept at first glance. However, different studies suggest contradictory findings regarding their effectiveness. On the one hand, researchers criticise that most current password meters just tell end-users if a password is weak without giving reasons (Ur et al. 2016). Renaud et al. (2017) evaluated different password nudges in an online experiment (N=1.273). Among others, they tested a dynamic password strength meter showing where on the x-axis the users' password is located. Similar to common meters, it showed only aggregated information, not revealing details about concrete dimensions. The tested nudges did not affect password strength. On the other hand, different studies have found that meters can yield stronger passwords. In an experiment with several meters, Egelman et al. (2013) found that meters can lead to stronger passwords when forcing users to change passwords on important accounts. As a conclusion, black-box-based password meters have in some cases proven to be effective.

Additionally to the common meter, other approaches enrich them with dynamic checklists, showing disaggregated information on password strength. For instance, those approaches highlight dimensions in a list (e.g. if the password contains a digit) to offer guidance to the user. While those visualisations can sometimes provide helpful feedback, in many cases they are very text-intensive and hard to capture. Ur et al. (2017) recently developed a password meter using neural networks and several heuristics to give detailed data-driven feedback. Due to the detailed text feedback, the approach can be considered white-box. They found that the new concept leads to more secure passwords than a common password meter bar while passwords are no less memorable (Ur et al. 2017). While this approach already focuses on the promising concept of giving detailed data-driven text feedback, additionally to the text, the output is a typical meter aggregating information.

Although enriched visualisations run the risk of being overwhelming, Renaud et al. (2017) suggest that offering guidance on how to achieve stronger passwords might be more effective than simply increasing awareness. Hence, they recommend using richer quantification measures for password strength in future works. A promising way to enrich the visualisation of password strength is by displaying a comprehensive picture of relevant information in a whitebox-based multidimensional visualisation. Kwon and Lee (2016), Santos, Haimes, and Lian (2007) and Yu and Liao (2016) suggest parallel coordinates to visualise multidimensional data, as it is an effective way to display information in a wide range of contexts. They found that although they are considered to be an unusual representation of information, participants of their study did not feel overwhelmed when using a simple dataset (Kwon and Lee 2016). Furthermore, they stated that the results could be generalised to other multidimensional visualisations like radar charts (Kwon and Lee 2016). Radar charts (also called star coordinate, spider chart or polar chart) are a common visualisation technique for disaggregation of data. They are also likely to have motivating effects due to its usage in games such as Pokemon.

2.3. Research gap

While nudging is a widely used and long-established instrument in many contexts, it has only recently emerged to be relevant in cybersecurity as well. Several works have focused on nudging towards awareness in privacy (e.g. Acquisti 2009; Balebako and Cranor 2014; Wang et al. 2014; Balebako et al. 2015; Saad and Khan 2016; Zhang and Xu 2016; Kaushal et al. 2017; Micallef et al. 2017; Wisniewski, Knijnenburg, and Lipford 2017; Alemany, Alberola, and García-Fornes 2019; Kroll and Stieglitz 2019; Zimmerman et al. 2019). Other research pays attention to nudging in the context of security (e.g. Ur et al. 2012; Jeske, Coventry, and Briggs 2014a, 2014b; Turland et al. 2015; Acquisti et al. 2017; Renaud et al. 2017; Renaud and Zimmerman 2017; Ur et al. 2017; Kankane, Dirusso, and Buckley 2018; Renaud and Zimmermann 2018b, 2018a; Jansen and van Schaik 2019; Peer et al. 2019; Van Bavel et al.

2019; Story et al. 2020). However, current research suggests that showing one-size-fits-all nudges does not lead to the desired outcomes. Hence, some studies address the potential of showing personalised nudges regarding individual characteristics. To our knowledge, personalised nudges in cybersecurity addressing preferences for white-box visualisation of information are largely missing.

To facilitate personalisation, the pool of effective nudges in cybersecurity has to be extended and adapted to distinct user groups. So far, the majority of nudges are blackbox-based, showing aggregated information. We suggest that those nudges are suitable for individuals who tend to feel overwhelmed by comprehensive information and are less likely to feel reactance when not knowing the underlying reasons. Furthermore, we suggest that it remains important to address a distinct group of individuals: users that make decisions based on disaggregated white-box information. In order to extend the pool of nudges in cybersecurity accordingly and therefore facilitate personalisation, we suggest to thoroughly evaluate visualisations of disaggregated information as whitebox-based nudges. In our study, we focus on an evaluation representative for the German population, providing an opportunity for comparability with other countries in future works.

3. Research design

The objective of this article is to address the lack of findings for whitebox-based multidimensional visualisations as nudges in cybersecurity while contributing to the emerging trend of personalisation. We aim to answer our research questions as described in the introduction. We address the cybersecurity context of nudging users towards creating stronger passwords which can be considered promising to provide better guidance for end-users (Zimmermann and Gerber 2020). Our intention is to evaluate two whitebox-based multidimensional visualisations. Following the idea of customised nudges, we thereby intend to provide specific users with the most suitable nudge.

In contrast to most related approaches (e.g. Egelman et al. 2013; Jeske, Coventry, and Briggs 2014b; Renaud et al. 2017; Peer et al. 2019), we aim to address multiple dimensions of password strength in a disaggregated presentation. Therefore, we first searched literature for dimensions of password strength that are suitable to assist in creating stronger passwords while being consistent with the white-box concept. Furthermore, we investigated which visualisation techniques are considered promising to effectively display disaggregated data. We found that radar charts and parallel coordinates are successfully used in different contexts to represent multiple dimensions. For instance, radar charts are commonly used in role-playing games to visualise a characters' strengths and weaknesses. We evaluated if both visualisation techniques are appropriate to display password strength in a transparent and comprehensible way. Therefore, we conducted a think-aloud prestudy. In a user-centred design process, we iteratively adapted the visualisations to the needs of our participants and the password creation context. Additionally, we conducted a representative online evaluation (N=1.012) to find out if our nudge significantly influences behaviour in password creation. For the experiment design, we implemented a dynamic radar chart and a commonly used password meter based on the zxcvbn.js calculation (Wheeler 2016) on a website which dynamically adapt to the password input. The *zxcvbn-score* is a low-budget password strength estimation which can reach a value between 0 and 4 to indicate if the cracking time is less than 10^2, 10^4, 10^6, 10^8 s or infinity (Wheeler 2016). For the interpretation of our results, it is crucial to notice that the complexity score of 4 is not further differentiable, resulting in a capped representation of password strength. Moreover, we integrated a survey on decision-making and information processing styles to segment user groups. Finally, we asked our participants to give an insight into their general attitude towards nudging in cybersecurity.

4. Think-aloud prestudy on n udging-types

We conducted a prestudy (N=20) in individual think-aloud sessions to investigate the comprehensibility of two different multidimensional visualisations (radar chart and parallel coordinates) in the context of password strength. Focusing on the white-box idea, we intended to evaluate if both visualisations can appropriately inform about the multiple dimensions of password strength in a comprehensible and transparent way. Comprehensibility is an important condition for being a successful whitebox-based nudge, as users need to understand how password strength was calculated and what to do to improve their password. Using low-fidelity mockups, we focused on comprehensibility before evaluating the actual nudging effect in our main study. Moreover, we intended to find advantages and disadvantages in comparison to the commonly used password meter. Thus we aimed to eliminate unsuitable visualisations at an early stage. You can find an example of all visualisations in Figure 3 and an image of the setting in Figure 4.

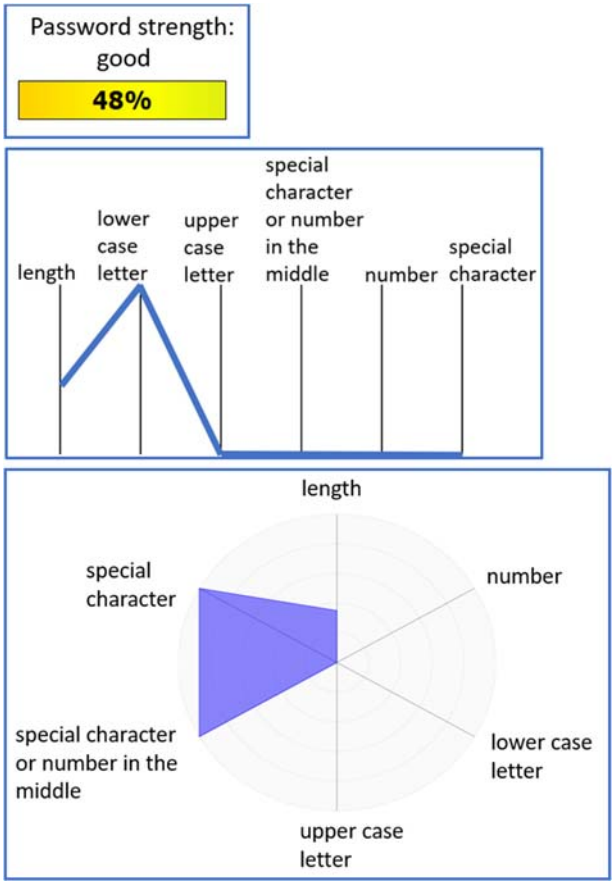

Figure 3. Examples of the visualisations used during our prestudy: password meter (top), parallel coordinates (middle), radar chart (bottom).

4.1. Method

We conducted a think-aloud study with 20 participants (11 female, 9 male) aged 24–35. Twelve test subjects stated to be university students, eight were employees or other. The average duration of one session was 12 min. Following the suggestions of Fonteyn, Kuipers, and Grobe (1993), individual sessions were conducted in a quiet setting. The sessions were audiotaped and a screen video was recorded using the Xbox DVR-tool. We asked our participants to explain their thoughts aloud while performing tasks. The participants were informed that there is no right or wrong in carrying out the given activities. Thus we intended to gain an insight into the thoughts of our participants concerning the following aspects: What did the participants like / dislike about the visualisations? Are the visualisations transparent? Did the participant feel overwhelmed? The investigator interfered only to remind the participant to keep thinking aloud if necessary.

Figure 4. Setting of the prestudy.

The method was conducted with three different visualisations of password strength on a computer screen which were shown in random order. We consecutively showed screenshots of a password and an associated visualisation of its strength, changing the visualisations multiple times. To each participant, we consecutively showed a common password meter, a radar chart and parallel coordinates. As the radar chart was used in our subsequent representative evaluation as well, please consider its detailed descriptions in Section 5.1. The participants were asked to answer questions concerning comprehension and preferences of the visualisations (e.g. 'Would you consider the given password strong?', 'Which visualisation do you prefer?'). In contrast to our following representative evaluation, we presented static nudges that did not encourage interaction. We selected the displayed passwords from the dataset of 32 million passwords that were leaked from the software company RockYou (Vance 2010). That dataset was already used in a similar context by Ur et al. (2016).

In accordance with the standard proceedings, we transcribed each session selectively, focusing on relevant information (Fonteyn, Kuipers, and Grobe 1993; Van Someren, Barnard, and Sandberg 1994). Those include both the instructions of the investigator and the thinking out loud of the test subject, but also '*unusual silences*' (Van Someren, Barnard, and Sandberg 1994) and information we gained from the screen video. Although the test subjects worked with static images of nudges on the screen, the video material provided some interesting information (e.g. pointing with the mouse). We assigned the information to their associated questions and tasks. In order to structure our findings, we deductively clustered the gained data thematically afterward. We assigned each relevant content to our six clusters (visual appearance, white-box characteristic, internal comparability, disaggregation, combination of radar chart and password meter, other) by considering the context and looking for keywords (e.g. 'colour scale').

For each cluster, we then drew a conclusion to consider in our next steps of the iterative evaluation process.

4.2. Results

We gathered the most important results by forming two main categories. The first category (characteristics of the nudges) refers to the clusters 'visual appearance', 'white-box characteristic', 'disaggregation' and 'internal comparability'. The second category (combination of radar chart and password meter) concerns the idea of combining features from both visualisations.

4.2.1. Characteristics of the nudges

When asking our participants about the visual appearance of the password meter, they agreed about the pleasant simplicity (e.g. *'It is pleasantly simple. (…) It reduces [the password strength] to just one number.'* (E1 #8:16)). Moreover, our participants liked the concept of showing traffic light colours. All participants highlighted their familiarity with the password meter. However, 17 out of 20 agreed that they missed feedback on what to do to improve their password. While some participants stated they were familiar with meters that additionally show a checklist, they still missed dynamic feedback concerning the actual input.

Nineteen out of 20 explicitly said they liked the disaggregated appearance of the radar chart and all were able to interpret the visualisation concerning password strength effortlessly. Many highlighted the motivating effect of the surface of the radar chart, which becomes bigger when the password is stronger. For some, it was a familiar and positively loaded appearance, as they knew it from video games (e.g. *'[That one] is really cool because I know it from video games and pen and paper '* (E3 #12:13)). When asked about the white-box characteristics of the radar chart, all test subjects could easily comprehend what was necessary to improve the password strength due to the disaggregated visualisation of dimensions (e.g. *'Here I have several links on what to improve (lists the dimensions) to enlarge the size of the surface '* (E2 #2:33)). Based on the coloured surface, the participants could understand which aspects of the password needed adjustment. For instance, they were able to transfer imbalances of the strength dimensions of a concrete password to the lopsided coloured surface. In the course of the evaluation of the radar chart, we included a warning when a password had occurred in a data breach as we considered it an important dimension of password strength. However, that characteristic of a password is the only dimension we did not map within the radar chart but next to it as it differs significantly from the other dimensions, not providing a clear implication for password improvement. While the warning was overall considered useful, some participants got confused when displaying both simultaneously, the radar chart and the warning. They were not sure if the size of the surface still mattered when the warning occurred. Our participants suggested to not display the radar chart when the warning appears. As the radar chart appeared to be the most promising novel nudge in our evaluation, we asked our participants to sort five entities of the radar chart nudge by password strength. Thereby, we intended to further look into the comparability of different strengths using the radar chart. All participants were able to correctly sort the five visualisations and hesitated only when two coloured surfaces had a very similar size.

As a third nudge, we evaluated the parallel coordinates. In contrast to the password meter and the radar chart, 19 out of 20 were confused by the visual appearance and struggled to interpret the visualisation. The participants had to think about it for a few moments before they were able to tell how strong a password was, which was accompanied by our third nudge. One participant specifically said that he expected the dimensions to be ordered according to their importance. However, there was no order intended from our side. That problem did not occur for the radar chart due to its circular appearance.

4.2.2. Combination of radar chart and password meter

When comparing our three nudges, some participants suggested combining features of the password meter and the radar chart. Since we had considered that idea to be promising, we had prepared such a visualisation and asked our participants for feedback. Our test subjects especially liked the simplicity of the traffic light colours of the password meter. However, they disliked that there was no convenient feedback on what dimensions of the password needed to change to become stronger. When showing a radar chart with the surface in traffic light colours, 17 out of 20 considered it a good combination and could imagine it to provide better assistance than each nudge individually. However, some test subjects were unsure about where to draw the line between the different colours.

4.2.3. Implications

Considering our findings of the think-aloud prestudy, we decided to eliminate the parallel coordinates from further evaluation. It turned out to be confusing and unfitting in our specific context of password strength, for instance because it suggests to sort dimensions by their importance from left to right which felt not

Table 1. Main characteristics of our approach.

Characteristic	Description
(1) Disaggregation	Contrary to a common password meter, we split up the visualisation in multiple relevant dimensions. The nudge displays the disaggregated information in a visualisation that is suitable for multiple dimensions. We chose the radar chart as a visualisation that has shown to be easily comprehensible for most participants of our prestudy.
(2) Whitebox-based and hybrid	Due to the disaggregation, the user can comprehend the feedback of the nudge for all implemented dimensions. It is clear what to do to receive positive feedback. All steps from the user input and the processing of the input to the output are transparent. Following the concept of Renaud and Zimmerman (2017), the nudge can be considered hybrid due to its enriched nature while displaying comprehensive information.
(3) Dynamic	The feedback of the nudge adapts dynamically to the input of the user.
(4) Personalisation	We intend to contribute to the pool of effective nudges by adding an alternative to black-box nudges for users that prefer transparent information over simplicity.

intuitive for many of our participants. However, besides the password meter, which has already proven to be suitable in several studies, the radar chart appears to be a promising visualisation technique for nudges in cybersecurity as well. Hence, in our further evaluation, we will focus on the two nudges, radar chart and password meter.

5. Representative evaluation

As the radar chart appeared to be promising to visualise multiple dimensions during our prestudy, we aimed to gain deeper insights into its potential as a nudge in comparison to a simple password meter. Therefore, we conducted a representative online evaluation with an integrated experiment. In Section 5.1, we present the design of the radar chart as a potential nudge for stronger passwords. The method of evaluation will be described in detail in Section 5.2. Furthermore, we suggest the findings of our representative study in Section 5.3.

5.1. Design of a White-box nudging mechanism

Based on the empirical findings of our prestudy, we designed a radar chart as a novel nudge displaying whitebox-based critical information in a multidimensional visualisation. The main characteristics of our approach are presented in Table 1.

To evaluate our approach, we chose the context of password creation as users still often fail to create safe passwords while preferring them as authentication scheme (Zimmermann and Gerber 2020). Aiming to display multiple aspects of password security, we identified six dimensions from related research that are suitable to evaluate our white-box approach in a first comprehensive step. It is crucial that the dimensions are transparent and easily comprehensible for the user. Therefore we focused on the following dimensions:

- password length (Yan et al. 2004; Veras, Thorpe, and Collins 2012; Ur et al. 2015, 2016)
- contains at least one number (Yan et al. 2004; Jakobsson and Dhiman 2013; Veras, Collins, and Thorpe 2014; Ur et al. 2015, 2016)
- contains at least one lowercase letter (Ur et al. 2015, 2016)
- contains at least one uppercase letter (Ur et al. 2015, 2016)
- contains at least one special character (Yan et al. 2004; Jakobsson and Dhiman 2013; Veras, Collins, and Thorpe 2014; Ur et al. 2015, 2016)
- contains at least one special character or number in the middle of the password (Ur et al. 2015, 2016)
- password has previously appeared in a data breach (https://haveibeenpwned.com).

We are aware of the fact that recommending complex passwords has been ruled out due to usability conflicts (Grassi et al. 2017; Tan et al. 2020). However, to evaluate the multidimensional visualisation as a nudge, it is necessary to test for the effects on the individual dimensions. Also, focusing on password length alone does not always result in better results as many users still tend to prefer shorter passwords. In those cases, it is beneficial to include other dimensions of password strength, for instance those that we included for the radar chart. Hence, the context of password creation using disaggregated information can be considered suitable for initial investigations.

While most other nudging approaches display information about password security in an aggregated form, we transparently display all identified dimensions. This idea of showing more detailed information was already implemented as a promising nudge by Ur et al. (2017). As other studies have shown (e.g. Kwon and Lee 2016) and as our prestudy has confirmed, radar charts are a suitable visualisation technique for disaggregated data. Moreover, radar charts are commonly used in the context of role-playing games to visualise a characters' strengths and weaknesses, suggesting a pleasant and familiar appearance for some user groups. Hence, there are reasonable grounds to evaluate the suitability

of radar charts in our context. Each identified dimension of password security despite the last one is mapped to a coordinate of the radar chart. To avoid misreadings and to encourage interpretations of the surface, we decided to arrange the dimensions in an ascending order regarding their respective values. The appearance in a data breach results in not displaying the chart at all, instead showing a warning message with the link to the source. A valuable feature of the radar chart is the mapping of information on the size of the surface. In our context, that means: the bigger the surface of the visualisation the stronger the password. It becomes immediately visible if the password has to be improved in a specific dimension. As length can be considered more important for password strength than the other dimensions, it is treated slightly differently. When the password has reached a specific length (e.g. 20 characters), the colour of the surface will be green, independently of the other dimensions. That suggests an acceptable password strength. In all other cases, the colour of the surface remains blue. You can find the design of our radar chart in Figure 5.

We implemented the nudge in a React App for the evaluation using *nivo*, a library integrating *React* and *d3*. To check the input passwords for appearance in data breaches, we used the API of *Have I been pwned* (https://haveibeenpwned.com). The visualisation dynamically adapts to the password input after clicking a button. Hence, the user receives individual feedback on the security of each password, he or she wants to have checked.

5.2. Method

To evaluate the potential of our radar chart as a nudge in comparison to a simple password meter, we conducted a representative survey with an integrated online experiment in Germany, including 1.012 participants. A pretest with 10 participants was performed beforehand to identify flaws and lack of clarity. We gathered our participants using the panel provider *Respondi*. Therefore, each participant was paid a small allowance (*e*1). Repeated participation was excluded and reviewed before the analysis. We included only participants that correctly answered all attention-check questions (e.g. 'Please select answer option number three.'). From an overall number of 1.635, 623 were immediately excluded from further participation due to failing the attention check questions or not fulfilling survey requirements (e.g. using a mobile phone). Hence, the total number of valid participants for the analysis was 1012.

Figure 5. Design of our radar chart.

The study was structured in the following subparts: (1) screening survey with demographic data, (2) segmentation of user groups using the *General Decision Making Style* scale (Scott and Bruce 1995) (subscales *rational* and *dependent*) and the *Rationale-Experiential Inventory* (Pacini and Epstein 1999) (subscales *rational ability* and *rational engagement*), (3) experiment and short usability scale, (4) survey on attitudes towards nudging in IT security. You can find the survey instrument in the appendix.

5.2.1. Segmentation of user groups

Since we aim to identify if specific user groups prefer our radar chart over a simple password meter, we needed to segment users concerning relevant characteristics. Therefore we chose to apply the psychometric *General Decision Making Style* scale (GDMS) and the *Rationale-Experiential Inventory* (REI) (Scott and Bruce 1995; Pacini and Epstein 1999). The GDMS scale measures how individuals make decisions. In a related study by Peer et al. (2019), the scale was already successfully used to personalise nudges. For our context of displaying white-box information in a radar chart, the subscales *rational* and *dependent* are particularly relevant. An individual with a high value in the rational-subscale makes decisions logically and systematically. An individual with a high value in the dependent-subscale prefers to rely on advice of other people and likes to be nudged towards good decisions (Scott and Bruce 1995). The REI scale measures preferences for information processing. For our segmentation of user groups, we chose the subscales *rational ability* (ability to think logically and analytically) and *rational engagement* (reliance on and enjoyment of thinking in an analytic, logical manner), as they appear to be promising indicators in the context of visualisation where analytical thinking is encouraged. All subscales are measured with a five-point Likert scale and can reach a total score from 1 to 5. Averaging the chosen subscales from GDMS and REI in a segmentation scale, we suggest the following hypotheses for user segmentation.

We assume that individuals with a high REI_ability and GDMS_rational score (between 3.0 and 5.0) and a low GDMS_dependent score (from 1.0 to 2.9) are more likely to be nudged by a white-box visualisation than by an aggregated password meter. On the other hand, we assume that individuals with a low REI_ability and GDMS_rational score and a high GDMS_dependent score prefer simple password meters, where thinking thoroughly is neither necessary nor helpful.

5.2.2. Experiment

After the segmentation of user groups, the participants were forwarded to a short online experiment. Investigating the ecological validity of password studies, Fahl et al. (2013) found that more than two-thirds of all participants in a role-playing scenario create passwords that mirror their real-world behaviour. Thus we decided to let our participants role-play to change their password for an email account. All participants were asked to create a password and type it in the password field of our interface. However, that interface differed among the test conditions. We randomly assigned our participants to one of three test conditions (see Figure 6). After data cleansing, 343 were assigned to condition 1, 325 to condition 2 and 344 to condition 3. The first condition (blue box in Figure 6) contains only the password field where the participants were asked to create a new password, not presenting a nudge. The second and third conditions additionally displayed a visual nudge, namely the password meter or the radar chart (see Figure 7). We chose between those nudges according to the determined results of the segmentation score. In the second condition (red box in Figure 6), we presented the nudge that was unfitting according to our hypotheses, namely a password meter when the score value of GDMS and REI suggested the radar chart, and a radar chart when the score value suggested the password meter. In the third condition (green box in Figure 6), we provided the nudge that was fitting best with regard to the scales' result. In conditions 2 and 3, the output of the nudge adapted dynamically according to the password input when clicking the button *'How strong is my password?'*. During the pretest, participants were partly confused by the dynamically changing visualisations during typing. Hence, for our representative experiment we decided to update the meter and the radar chart not while typing in the password, but after clicking the button. It is however important to consider that the interaction comes with an additional user effort. All participants could change the password as often as they liked until they confirmed by clicking a button. We would like to point out that we did not collect data on how many times participants clicked on the button and how that might have differed between test conditions or nudges. Passwords could not be submitted without seeing or updating the nudge. When confirmed, independently of the test condition, several password characteristics (length, amount of lower cases, upper cases, digits, special characters and special characters or digits in the inner of the password, pwned-boolean and zxcvbn-score) were locally calculated and stored in a JSON file for later analysis. Passwords themselves were stored at no time and were hashed before transmitting to the *Have I been pwned*-calculation to protect our participants. After the experiment, participants of condition 2 and 3 were forwarded to a slightly adapted System Usability Scale (SUS) (Brooke 1996).

5.2.3. Survey on the attitude towards nudging in cybersecurity

While there is a lot of research on how people perceive nudges in healthcare or advertisements, little has been evaluated regarding the attitude towards nudges in IT security. Therefore, we conducted a survey, focusing especially on gaining insights into the attitude towards white-box nudging.

5.2.4. Analysis

To examine the obtained data of the online evaluation, we used *R* and *RQDA* for data preparation and statistical

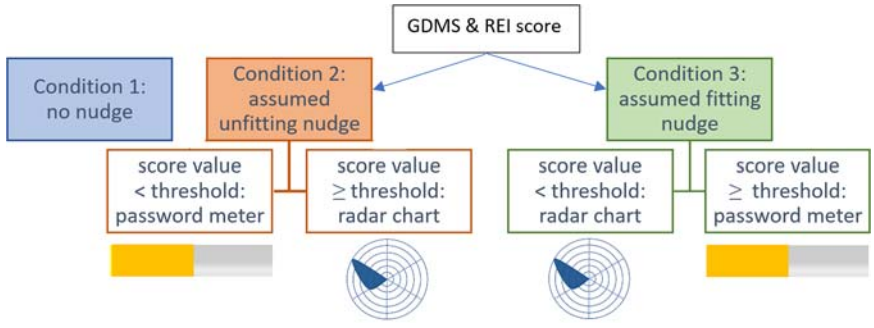

Figure 6. Test conditions of the experiment.

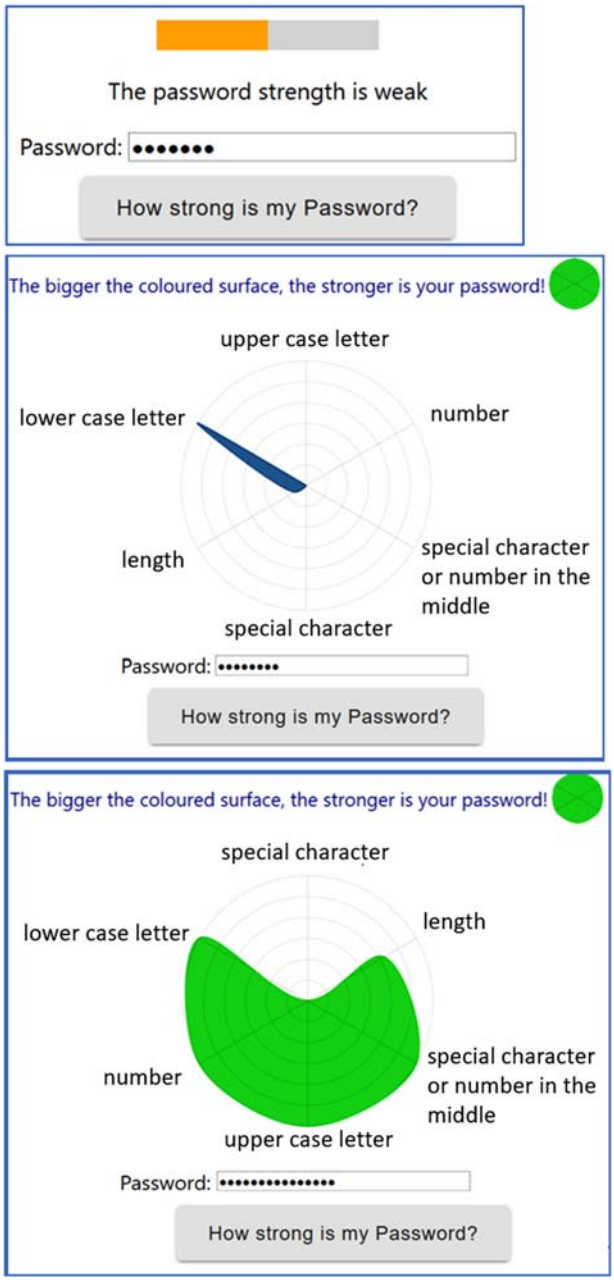

Figure 7. The interface of our main study with a simple password meter (top) and our novel nudge referring to a weak password (middle) and a strong password (bottom).

analysis and *Tableau* for creating data figures. We calculated the basic frequencies for each item. Also, we determined significant differences in gender, age, education, income and place of residence (federal states) regarding our sample and the German population by applying the χ^2-test of independence. To identify significant differences between groups concerning our hypotheses, we performed Student's *t*-tests, adjusting the determined *p*-values with Bonferroni correction to encounter the problem of multiple comparisons. When the samples were of unequal size, we applied the Welch's unequal variances *t*-test instead. We set the significance level to $\alpha = 0.05$. To take into account multiple dependent variables, we performed multivariate analyses of variance (*MANOVA*). Beforehand we tested for homogeneity of variance as necessary condition, using the *Levene's test*. For all examined variables, the Levene's test was not significant, providing the necessary conditions for *MANOVA* testing. To identify which mean values were responsible for significant *MANOVA* results, we subsequently performed post-hoc tests. Therefore, we used the *Tukey's HSD*-test which is suitable for multivariate analyses of variance.

5.2.5. Characteristics of survey participants

The conducted online experiment is representative of the German population according to gender, age from 18 to 74, education and income. Further, the participants show little difference to the German population regarding federal states, as we ensured a wide spread and an approximate proportional distribution. We tested our sample for its representative nature using the χ^2-test. Regarding gender, the χ^2-test reveals no significant differences to the German population ($\chi^2(df = 1) = 0.32, p = .5715$). Our sample consists of 48.72% female and 51.28% male participants (Germany: 50.43% female and 49.57% male). The participants age was gathered in groups of 18–29 years, 30–39 years, 40–49 years, 50–59 years and 60–74 years. All groups are representative of the target population ($\chi^2(df = 4) = 3.54, p = .4720$). We asked our participants to assign their income to one of three groups (under *e*2000, *e*2000 to *e*4000, above *e* 4000). All groups are represented proportionally ($\chi^2(df = 2) = 3.35, p = .1876$). In order to collect results representing the German population concerning education, we gathered information on the highest level of education in three groups (without a school diploma/certificate of secondary education ('Hauptschulabschluss'), general certificate of secondary education ('mittlere Reife'), qualification for university entrance ('Abitur')/university degree). According to the χ^2-test, all groups are represented proportionally to the German population. Moreover, all 16 German federal states are represented in an approximate proportion to the German population. Additionally, we asked our participants if they liked to play role-playing games since that characteristic was assumed to be important for later analysis. 46.34% stated to like playing role-playing games, 51.28% stated to not like playing them.

5.2.6. User segmentation

To evaluate our novel nudge in a personalised setting, we segmented our user groups regarding decision-

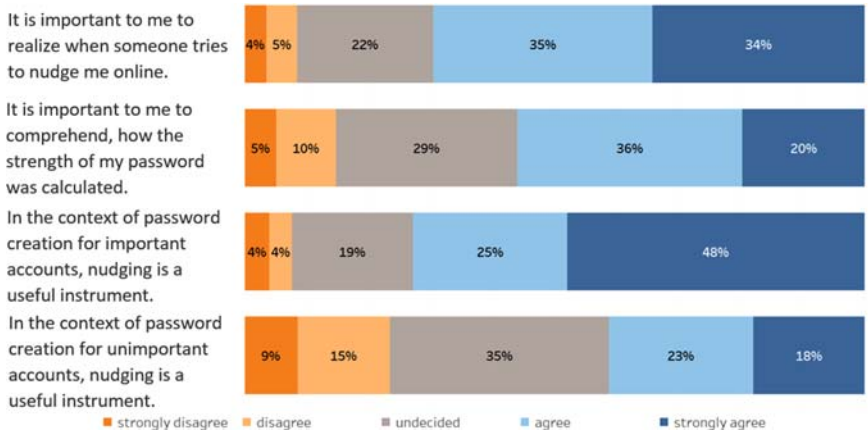

Figure 8. Results on attitudes towards nudging in cybersecurity.

making and information-processing style as described in Section 5. Averaging four subscales in total, the participants were able to reach a segmentation score from 1.0 to 5.0. The mean score of our sample was 3.54 (MIN = 1.14, MAX = 4.80, $SD = 0.49$). Following our initial hypothesis, we assumed that participants with a score of less than 3.0 would prefer a simple password meter over the comprehensive radar chart. Furthermore, we assumed that participants with a score of 3.0 and higher would prefer the radar chart due to its white-box nature. 10.87% of our participants scored less than 3.0 and 89.13% scored 3.0 or more.

5.3. Results

We present our findings of the representative online study by starting with general questions concerning the attitude towards white-box nudging in cybersecurity. Afterwards, we suggest the results of our novel whitebox-based nudge – the radar chart – in comparison to the commonly used password meter, where we identified differences regarding nudging efficacy. Also, we identified group differences between players of role-playing games and non-players.

5.3.1. Attitude towards white-box nudging in cybersecurity

To gain a representative insight into attitudes on and perceptions of white-box nudging in cybersecurity, we correspondingly included several questions in our online study. You can see an excerpt of our results in Figure 8. 69.6% agreed it was important to them to realise if someone tried to nudge them online (35.4% agreed, 34.2% strongly agreed). When asked more specifically, 56.0% agreed, it was important to them to comprehend how the assessment of password strength was calculated. Furthermore, we asked our participants if they thought nudging was useful in the context of password creation. 72.9% agreed to nudging being a useful instrument for password creation regarding important accounts (24.9% agreed, 48.0% strongly agreed). Regarding unimportant accounts, 41.2% still agreed.

Moreover, we asked our participants in free-text questions if they thought nudging in the context of cybersecurity was useful or if they perceived risks. We deliberately set the question as optional to prevent poor quality of answers by participants that do not feel capable answering. Therefore, of 1012 participants we had to exclude 523 because they decided not to answer the question or explicitly stated to be too undecided to give a specific input. Using *RQDA* independently by two researchers, the remaining 489 answers were deductively assigned to one of four clusters (rating nudges as useful, undecided and thoroughly evaluated answers, rating nudges as inappropriate, comments on potential risks) by looking for keywords (e.g. 'useful' or ' patronised') and thoroughly analysing the answers. Clusters were assigned in a two-step approach. The answers were first grouped roughly into unambiguously positive or negative answers as well as contents that needed further inspection before making a decision. Then, the answers were assigned to one of the four final clusters while concentrating particularly on the initially ambiguous contents. After the independent clustering by two people, the inter-rater reliability was calculated, resulting in a substantial level of reliability with a Cohen's kappa coefficient of $\kappa = 0.755$ (Cohen 1960).

We found that about half of the participants ($rater_1 = 59\%$, $rater_2 = 56\%$) explicitly commented positively on nudging in cybersecurity. Several participants highlighted the importance of transparency (e.g. *'As long as the final decision is up to the user, [the*

nudge] can improve IT security and that is partly urgently necessary. It has to be completely transparent, how the nudges work and make decisions' (E157)). Also, around 15% ($rater_1 = 14\%$, $rater_2 = 16\%$) were undecided if nudging in cybersecurity is useful or not while thoroughly evaluating advantages and risks within their answer (e.g. *'On the one hand it is useful to have nudges when helping unexperienced people by nudging them in a safe direction. On the other hand, criminals / hackers can use nudges to manipulate people and get access to their data'* (E577)). While many participants commented positively or balanced, some ($rater_1 = 13\%$, $rater_2 = 8\%$) explicitly stated they did not agree with nudging being useful. In equal parts, reasons were first, the perception of nudges as superfluous but not dangerous and second, fearing the risk of paternalism, manipulation, censorship or data exposure. Moreover, some ($rater_1 = 14\%$, $rater_2 = 20\%$) did not answer if they rated nudging as useful but only commented on potential risks: 24 participants stated to see no risks, 8 participants only named advantages of nudging and 75 participants only enumerated potential risks.

Both radar chart and password meter have a slight influence on the general attitude towards nudging in cybersecurity. Participants were slightly less likely to rate nudging in cybersecurity as dangerous ($M_{nudge} = 2.24$, $SD_{nudge} = 1.0$, $M_{none} = 2.73$, $SD_{none} = 1.1$; $t(663.63) = -6.96$, $p < .0001$), patronising ($M_{nudge} = 2.48$, $SD_{nudge} = 1.1$, $M_{none} = 2.84$, $SD_{none} = 1.2$; $t(642.96) = -4.70$, $p < .0001$) or superfluous ($M_{nudge} = 2.25$, $SD_{nudge} = 1.1$, $M_{none} = 2.65$, $SD_{none} = 1.2$; $t(625.22) = -5.25$, $p < .0001$) when they were assigned to the password meter or radar chart during the previous online experiment. In contrast, participants that carried out the experiment in the control group without a nudge were slightly more likely to rate nudging in those negative ways. According to our results, it made no difference for the attitude towards nudging if the radar chart or the password meter was shown, as long as one of these nudges was assigned.

5.3.2. Differences in nudging efficacy

Evaluating the efficacy of our tested nudges is a central aspect. Short-term efficacy was tested as an initial step to gain insights into the nudges' potential. We evaluated if the participants were immediately showing a different behaviour concerning password creation when given a nudge. Hence, we examined all seven dimensions included in our white-box visualisation as well as the zxcvbn-score. You can find an overview of the mean values in comparison in Table 2. To prevent wrong conclusions on group differences, we first checked for general differences in password creation concerning the segmentation score. We found that participants with a low segmentation score tended to create slightly weaker passwords than participants with a high score. The effect was existent also when looking at the subscales separately (see Tables A2 and A3). Furthermore, we checked if the time effort between participants that interacted with the radar chart differed from those who interacted with the password meter. We found no significant differences of time effort during the process of password creation ($t(433.42) = 0.62$, $p = .5364$).

Table 2. Average password characteristics for different nudges.

Dimension	None
length	M=10.34,SD=4.2
# digits	M=2.52,SD=2.3
# lowercase letters	M=6.34,SD=4.3
# uppercase letters	M=1.05,SD=1.4
# special characters	M=0.43,SD=0.9
# spec. char./digits in middle	M=1.06,SD=1.1
zxcvbn-score	M=2.45,SD=1.3
Dimension	**Meter**
length	M=12.00,SD=4.9
# digits	M=2.97,SD=2.2
# lowercase letters	M=6.90,SD=4.9
# uppercase letters	M=1.40,SD=2.0
# special characters	M=0.68,SD=1.1
# spec. char./digits in middle	M=1.35,SD=1.0
zxcvbn-score	M=2.91,SD=1.2
Dimension	**Radar**
length	M=12.28,SD=5.4
# digits	M=2.97,SD=2.3
# lowercase letters	M=7.13,SD=4.9
# uppercase letters	M=1.33,SD=1.5
# special characters	M=0.83,SD=1.4
# spec. char./digits in middle	M=1.43,SD=1.2
zxcvbn-score	M=2.93,SD=1.2

To facilitate readability, detailed results of the MANOVA and post-hoc tests regarding nudging efficacy can be found in the Appendix in Tables A1–A3 instead of the text body. Mean values and standard deviations are presented in Table 2. Our novel nudge – the radar chart – was effective in nudging our participants towards stronger passwords regarding short-term effects in most evaluated dimensions. The nudge delivered significant differences in comparison to our control group without a nudge concerning password length, as well as slight differences concerning usage of digits, special characters and special characters or digits in the inner of the password. Further, the radar chart achieved a slightly better zxcvbn-score compared to not showing a nudge and participants were less likely to use a password that had appeared in a data breach ($t(657.34) = -4.77$, $p = .0002$). It is, however, crucial to consider that the effect on not appearing in a data breach was achieved due to the displayed warning.

Our results also confirm the findings of other research, namely that a password meter can positively affect password strength on short term as well.

Participants that were given a password meter showed significantly better results than participants in our control group regarding password length, as well as slightly better results in the appearance of digits, upper case letters, special characters and special characters or digits in the inner of the password. They achieved a slightly better zxcvbn-score and were less likely to choose a password that had previously appeared in a data breach ($t(662.47) = -3.84$, $p = .0098$). Hence on our initial investigation, password meter and radar chart appear to be similarly effective nudges regarding short-term effects. When looking more specifically into the results of the REI and GDMS subscores, we found that it made no difference for participants with a low (less than 3.0) or high (3.0 to 5.0) subscore regarding password strength if they were assigned to either the meter or the radar chart. Hence, participants that were assigned to the password meter created just as strong passwords as participants that were assigned to the radar chart, independently of their results in decision-making and information processing styles (e.g. Tukey's HSD test for radar chart vs. meter, low REI_ability-score and low gdms_rational-score regarding zxcvbn-score: p=.9908). To further evaluate if our suggested segmentation for showing personalised nudges worked, we compared the password strength of participants in condition 3 (showing the assumed suitable nudge) with random choice of nudges. We did not detect significant differences between those groups. Furthermore, we compared participants in condition 2 (showing the assumed unsuitable nudge) with participants in condition 3. Again, we did not find significant differences in all dimensions of password strength.

5.3.3. Differences in usability perception

The perceived usability of the password meter according to the slightly adapted SUS was M=71.57 of 100 points on average (SD=17.7). For the radar chart, the perceived usability was M=69.90 (SD=19.0). We applied a Student's t-test and found no significant differences for the usability of both nudges (p=.8978). Therefore, the usability of both visualisations can be considered good (Brooke 1996). Again, to prevent wrong conclusions, we investigated if participants with a low segmentation score rated the usability of password meter or radar chart differently than participants with a high segmentation score. However, no significant differences were detected (p=.5125). When comparing conditions 2 (assumed unsuitable nudge) and 3 (assumed suitable nudge), we found that participants with the suitably assigned nudge (high segmentation score) were significantly more likely to rate the radar chart more positively than participants with the unsuitably assigned nudge (low segmentation score). For instance, they stated significantly more often to feel confident using the radar chart ($t(33.72) = 3.329$, $p = .0021$) and rated the visualisation as easy to use ($t(32.97) = 2.818$, $p = .0081$). Surprisingly, participants with the unsuitable nudge (high segmentation score) were in addition more inclined to like to use the password meter frequently than participants in condition 3 ($t(44.43) = 4.904$, $p < .0001$).

5.3.4. Differences between gamers and n on-gamers

As the radar chart is commonly used in several role-playing games to visualise character strengths and weaknesses, we investigated if there is an interesting difference between players and non-players concerning our nudges.

Players and non-players generally did not differ in the investigated password strength dimensions (e.g. zxcvbn-score: p=.4516). Also, there were no significant differences regarding password strength detected between gamers that were assigned to the password meter and gamers that were assigned to the radar chart. In contrast to the results concerning nudging efficacy, we identified small but highly significant differences between gamers and non-gamers in respect of particular items of the radar chart's usability scale. Hence, there are tendencies other than groundbreaking effects. The detailed results are presented in Table 3.

For instance, on the Likert scale from 1 to 5, gamers were more likely to agree with the item *"I think that I would like to use this visualisation frequently"* than non-gamers. Moreover, gamers rated the radar chart as slightly easier to use than non-gamers and were slightly more likely to imagine it to be quick to learn

Table 3. Significant results of the usability scale for the radar chart between gamers and non-gamers.

SUS item	mean gamers	SD gamers	mean non-gamers	SD non-gamers	t-test result
'I think that I would like to use this visualisation frequently'	M=3.69	SD=1.1	M=3.18	SD=1.1	$t(318.71) = 4.29$, $p < .0001$
'I though the visualisation was easy to use'	M=3.95	SD=0.9	M=3.68	SD=1.1	$t(333.11) = 2.53$, $p = .0118$
'I would imagine that most people would learn to use this visualisation very quickly'	M=3.93	SD=0.9	M=3.63	SD=1.1	$t(334.96) = 2.78$, $p = .0054$
'I felt very confident using the visualisation'	M=3.85	SD=1.0	M=3.54	SD=1.1	$t(329.75) = 2.69$, $p = .0076$

for most people. Overall, gamers tended to be slightly more confident using the radar chart than non-gamers.

6. Discussion and conclusion

Recently, nudging in cybersecurity has emerged to be a relevant instrument. While personalisation is suggested as vital to obtain good results for individuals online, so far little has been evaluated in the context of cybersecurity (Peer et al. 2019). Furthermore, enriching nudges by showing white-box information in an appealing visualisation is a promising trend to address flaws of simple black-box nudges (e.g. reactance, missing comprehensibility) that were criticised for instance by Ur et al. (2016). Our study contributes to the research landscape of nudging in cybersecurity, addressing current trends such as personalisation and transparency. Hence, our scientific contribution is an initial evaluation of different white-box nudges (radar chart and parallel coordinates) in comparison to a simple black-box password meter in the context of password creation.

To answer our first research question (*'How can whitebox-based multidimensional visualisations provide an effective nudge towards better security decisions in password creation for specific user groups?'*), the findings reveal that the radar chart is a moderately effective nudge regarding short-term effects, displaying multiple dimensions in an appealing visualisation.

(1) The radar chart had a significant short-term effect on password strength. We found that it encouraged users create passwords that were slightly harder to crack, affecting for instance length and the appearance of digits, special characters and uppercase letters. It is, however, important to point out that the long-term effectiveness and evaluations in a real-world scenario are still pending and should be particularly considered to gain more realistic insights.
(2) Like the radar chart, the password meter resulted in stronger passwords, supporting the findings concerning nudging efficacy of Egelman et al. (2013).
(3) When comparing the short-term efficacy of the white-box radar chart and the black-box password meter in general, radar chart and password meter can be considered approximately equally effective nudges at first glance. Regarding time effort, we furthermore found no significant differences between the two nudges. Following the idea of white-box information, which was suggested for instance by Renaud et al. (2017), we assumed that comprehensive and transparent nudges were preferred over simple black-box nudges by certain user groups.
(4) Our evaluation reveals that users tend to prefer transparent nudges over black-box nudges. According to our survey, 56% want to comprehend how the assessment of password strength was calculated online. More generally, the majority of German citizens do not want to be nudged unconsciously in the context of cybersecurity. 70% stated they wanted to know if someone tried to nudge them online. However, the concept of nudging in cybersecurity was assessed positively by half of the participants (52%) or balanced by others (14%). Hence, we suggest expanding research on transparent nudges in cybersecurity and propose comparable studies on the attitude towards nudging in cybersecurity for other countries. While Peer et al. (2019) successfully used decision-making styles for personalisation, we took up that suggestion and complemented it with the information processing style. Surprisingly to us, the results of the questionnaires were behaving contrary to our expectations to personalise nudges. While we expected users with a low result in the REI_ability or GDMS_rational subscores to prefer the simple password meter, they did not create significantly stronger passwords than when assigned to the radar chart. Users with a high score performed equally regarding password strength when given the radar chart and the password meter as well.
(5) Thus personalisation via decision-making and information processing styles did not work for our experiment. Interestingly, we found a further indication for the assignment of either a black-box or a white-box nudge. Participants that like to play role-playing games rated the radar chart slightly better than non-gamers. For instance, they felt significantly more confident while using it. As stated before, the radar chart is commonly used to visualise a characters' strengths and weaknesses in role-playing games and likely provides a motivating effect to certain individuals. However, our investigations on that matter are not yet comprehensive enough to draw final conclusions.
(6) Hence, we suggest to expand research on gaming preferences as indicators for the successful personalisation of online nudges.

Moreover, we intended to answer our second research question: *'Do users feel overwhelmed by a more comprehensive and multidimensional visualisation of information in the context of security decisions in*

Table 4. Design implications for nudging in cybersecurity.

Challenges	Implications
(1) Facilitation of trust in and comprehensibility of nudges	To facilitate trust in nudges for cybersecurity, transparent nudges are preferred over black-box nudges. Hence, designers may uncover when end-users are nudged, give information about the intentions and benefits, and – where possible – provide white-box information on how the nudge is composed.
(2) Making use of familiar features for nudges	Not surprisingly, familiarity with visualisations plays an important role for the intuitive interpretation of nudges. Thus, providing user groups with a visualisation they are familiar with, can be favourable for usability perceptions. Combining novel visualisations with familiar features (e.g. traffic light colours) may, however, be sensible as well.
(3) Balance between information provisioning and simplicity of nudges	Designers of nudges may evaluate carefully if a disaggregated nudge is sensible for a specific context and a specific user group. While many end-users may prefer transparency, others may feel overwhelmed. Also, simple nudges may be generally more suitable for some specific contexts where multidimensional information is irrelevant.

password creation?' The parallel coordinates as a multidimensional visualisation appeared to be unsuitable for password strength. The participants were confused and struggled to interpret the visualisation. However, our participants were able to interpret the radar chart effortlessly. Even participants with a low score in information processing were successfully nudged towards stronger passwords by the radar chart in our short-term experiment. However, the effectiveness in a real-world scenario is still pending. Thus we suggest the radar chart as a potential research object in future works as a white-box visualisation for nudges in cybersecurity and as a potential contribution to the pool of effective nudges.

Summarising our main findings regarding the design of nudges in cybersecurity, we suggest three design implications. You can find them in Table 4.

7. Limitations and future work

While the study we report is a first step to extend the pool of effective nudges in cybersecurity for personalisation, it has also limitations.

(1) We acquired some of the results using a survey. Data collection through a survey is prone to social desirability biases as it is based on answers of individuals rather than the observation of actual behaviour. Also, the survey questions on attitudes were not based on a standardised and previously validated test. Therefore, other methods need to confirm that the questions measured attitudes adequately. However, our focus was to gain first representative insights into people's perceptions and, hence, the study provides valuable results for that matter.

(2) Moreover, 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 of the whitebox-based nudge by addressing the important metric of password memorability. While our work is a first step to investigate the potential of white-box 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 utilise other techniques of data collection in the future. Also, participants in our prestudy felt partly confused by the dynamic updates while typing in the password. Thus we decided to update both meter and radar chart only after clicking a button. As a conclusion, user effort may play a significant role in our evaluation and has to be considered prospectively. Hence, for future studies we plan to evaluate the effect of dynamically updating the nudges with a time delay while typing, to potentially minimise both confusion and user effort. For now, the evaluation can be considered a worthwhile initial step to gain insights into novel nudging visualisations, especially as the study was accompanied by qualitative investigations as well.

(3) Furthermore, the radar chart was tested against a commonly used black-box password meter. However, there is already another whitebox-based approach for password creation by Ur et al. (2017) which may be tested against in a next step to facilitate comparability regarding efficacy and usability. Both approaches, the radar chart and the meter by Ur et al., differ significantly in their visual representation of transparent information. Thus in a next step future works may gain interesting insights on advantages and disadvantages of the distinct nudges. We also suggest to evaluate other white-box visualisations or combinations and contexts. Nevertheless, we consider testing the radar chart against a blackbox-based password meter in a first step valuable, as it is still very commonly used online. Further, the long-term intention is to introduce a potential additional nudge to the pool of effective nudges to facilitate personalisation, rather than replacing existing nudges in general.

(4) Although the radar chart itself has proven to be a suitable visualisation for multidimensional information by several studies, it comes also with several limitations which may be investigated more intensively in future studies. For instance, to gain sensible surfaces, the dimensions have to be arranged in an ascending or descending order. As a result, when typing in multiple passwords consecutively, the user will be presented with radar charts that differ in the arrangement of their dimensions. During our pre-study, most participants were not confused, however, there may be a significant effect on usability which is still to be addressed. Also, the radar chart might be difficult to implement for mobile devices due to its size and the limited readability of dimensions in a compact UI. Furthermore, the radar charts suggests equal importance of dimensions, which is misleading in the context of password strength. We addressed that in a first step by treating length slightly differently than the other dimensions. However, future work may explore if manipulations of the distance between axes can be a more appropriate solution. Moreover, there are other interpretation issues concerning the size of the surface which may be considered in future research to address, for instance, if doubling the surface of the radar chart makes the user assume that the password strength was doubled as well. Including a warning when the password has appeared in a data breach separately from the radar chart comes with a bias towards password strength regarding that specific characteristic. Hence, to make implications for the radar chart as a stand-alone nudge concerning passwords appearing in a data breach, future studies should consider testing the radar chart without the warning or finding solutions to integrate it within the chart itself.

(5) Some differences, for instance regarding comparisons between gamers and non-gamers, cannot be considered large. Hence, they do not reveal groundbreaking differences, but still tendencies were identified that may be investigated more thoroughly in a next step. We suggest to investigate the potential of considering differences between people that play role-playing games and people that do not in future works, taking more specifically into account in which role-playing games radar charts are present.

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. We would like to thank Sebastian Linsner for his conceptual contributions prior to this study.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported 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.

ORCID

Katrin Hartwig http://orcid.org/0000-0003-4875-0110
Christian Reuter http://orcid.org/0000-0003-1920-038X

References

Abawajy, J. 2012. "User Preference of Cyber Security Awareness Delivery Methods." *Behaviour and Information Technology* 33: 237–248. doi:10.1080/0144929X.2012.708787.

Abdul, A., J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli. 2018. "Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda." In *Conference on Human Factors in Computing Systems -- Proceedings*, 1–18. doi:10.1145/3173574.3174156.

Acquisti, A. 2009. "Nudging Privacy -- The Behavioral Economics of Personal Information." *IEEE Security and Privacy* 7: 82–85. doi:10.1109/MSP.2009.163.

Acquisti, A., M. Sleeper, Y. Wang, S. Wilson, I. Adjerid, R. Balebako, L. Brandimarte, et al. 2017. "Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online." *ACM Computing Surveys (CSUR)* 50: 44. doi:10.1145/3054926.

Alemany, J., J. Alberola, and A. García-Fornes. 2019. "Enhancing the Privacy Risk Awareness of Teenagers in Online Social Networks Through Soft-paternalism Mechanisms." *International Journal of Human Computer Studies* 129: 27–40. doi:10.1016/j.ijhcs.2019.03.008.

Balebako, R., and L. Cranor. 2014. "Improving App Privacy: Nudging App Developers to Protect User Privacy." *IEEE Security and Privacy* 12: 55–58. doi:10.1109/MSP.2014.70.

Balebako, R., F. Schaub, I. Adjerid, A. Acquisti, and L. F. Cranor. 2015. "The Impact of Timing on the Salience of Smartphone App Privacy Notices." *SPSM 2015 -- Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices, co-located with: CCS 2015*, 63–74. doi:10.1145/2808117.2808119.

Bhuiyan, M. M., K. Vick, T. Mitra, K. Zhang, and M. A. Horning. 2018. "FeedReflect: A Tool for Nudging Users to Assess News Credibility on Twitter." In: *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, Association for Computing Machinery*, 205–208. doi:10.1145/3272973.3274056.

Biselli, T., and C. Reuter. 2021. "On the Relationship Between IT Privacy and Security Behavior: A Survey Among German Private Users." *16th International Conference on Wirtschaftsinformatik*, 1–17.

Boyce, M. W., K. M. Duma, L. J. Hettinger, T. B. Malone, D. P. Wilson, and J. Lockett-Reynolds. 2011. "Human Performance in Cybersecurity: A Research Agenda." *Proceedings of the Human Factors and Ergonomics Society*, 1115–1119. doi:10.1177/1071181311551233.

Brooke, J. 1996. "SUS -- A Quick and Dirty Usability Scale." *Usability Evaluation in Industry* 189: 4–7. www.TBIStaffTraining.info.

Chen, J., and A. Abouzied. 2016. "One LED is Enough: Catalyzing Face-to-face Interactions At Conferences with a Gentle Nudge." *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* 27: 172–183. doi:10.1145/2818048.2819969.

Cheng, H. F., R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, and H. Zhu. 2019. "Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders." *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems -- CHI '19*. doi:10.1145/3290605.3300789.

Cohen, J. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20: 37–46.

Dupree, J. L., R. Devries, D. M. Berry, and E. 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, 5228–5239. doi:10.1145/2858036.2858214.

Egelman, S., M. Harbach, and E. Peer. 2016. "Behavior Ever Follows Intention?: A Validation of the Security Behavior Intentions Scale (SeBIS)." *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5257–5261, ACM. doi:10.1145/2858036.2858265.

Egelman, S., and E. Peer. 2015. "The Myth of the Average User: Improving Privacy and Security Systems Through Individualization." *Proceedings of the 2015 New Security Paradigms Workshop*, 16–28. doi:10.1145/2841113.2841115. arXiv:arXiv:1508.06655v1.

Egelman, S., A. Sotirakopoulos, I. Muslukhov, K. Beznosov, and C. Herley. 2013. "Does my Password go up to Eleven? The Impact of Password Meters on Password Selection." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2379–2388. ACM. doi:10.1145/2470654.2481329.

Fahl, S., M. Harbach, Y. Acar, and M. Smith. 2013. "On the Ecological Validity of a Password Study." In: *Proceedings of the Ninth Symposium on Usable Privacy and Security*, 13. ACM.

Fonteyn, M., B. Kuipers, and S. Grobe. 1993. "A Description of Think Aloud Method and Protocol Analysis." *Qualitative Health Research* 3: 430–441.

Grassi, P. A., J. L. Fenton, E. M. Newton, R. A. Perlner, A. R. Regenscheid, W. E. Burr, J. P. Richer, et al. 2017. *NIST Special Publication 800–63B, Digital Identity Guidelines*, Technical Report. NIST. https://pages.nist.gov/800-63-3/sp800-63b.html.

Hartwig, K., and C. Reuter. 2019. "TrustyTweet: An Indicator-based Browser-Plugin to Assist Users in Dealing with Fake News on Twitter." In: *Proceedings of the International Conference on Wirtschaftsinformatik (WI)*, 1858–1869.

Hartwig, K., and C. Reuter. 2020. "Fake News Technisch Begegnen – Detektions- und Behandlungsansätze zur Unterstützung von NutzerInnen (in german)." In: *Wahrheit und Fake News im postfaktischen Zeitalter*, edited by P. Klimczak, T. Zoglauer. Springer.

Herbert, F., G. M. Schmidbauer-Wolf, and C. Reuter. 2020. "Differences in IT Security Behavior and Knowledge of Private Users in Germany." *Proceedings of the International Conference on Wirtschaftsinformatik (WI)*, 168–184. doi:10.30844/wi_2020_v3-herbert.

House, D., and M. K. Raja. 2019. "Phishing: Message Appraisal and the Exploration of fear and Self-Confidence." Behaviour and Information Technology. doi:10.1080/0144929X.2019.1657180.

Huh, J. H., H. Kim, S. S. V. P. Rayala, R. B. Bobba, and K. Beznosov. 2017. "I'm too Busy to Reset my LinkedIn Password." 387–391. doi:10.1145/3025453.3025788.

Jakobsson, M., and M. Dhiman. 2013. "The Benefits of Understanding Passwords." *Mobile Authentication*, 5–24. doi:10.1007/978-1-4614-4878-5_2.

Jansen, J., and P. van Schaik. 2019. "The Design and Evaluation of a Theory-based Intervention to Promote Security Behaviour Against Phishing." *International Journal of Human Computer Studies* 123: 40–55. doi:10.1016/j.ijhcs.2018.10.004.

Jansson, K., and R. Von Solms. 2013. "Phishing for Phishing Awareness." *Behaviour and Information Technology* 32: 584–593. doi:10.1080/0144929X.2011.632650.

Jeske, D., L. Coventry, and P. Briggs. 2014a. "Decision Justifications for Wireless Network Selection." *2014 Workshop on Socio-Technical Aspects in Security and Trust*, 1–7. IEEE. doi:10.1109/STAST.2014.9.

Jeske, D., L. Coventry, and P. Briggs. 2014b. "Nudging Whom How: IT Proficiency, Impulse Control and Secure Behaviour." *Proceedings of the CHI Workshop on Personalizing Behavior Change Technologies*, 1–4.

Kankane, S., C. Dirusso, and C. 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*, 1–6. ACM.

Kaufhold, M. A., N. Rupp, C. Reuter, and M. Habdank. 2020. "Mitigating Information Overload in Social Media During Conflicts and Crises: Design and Evaluation of a Cross-Platform Alerting System." *Behaviour & Information Technology (BIT)* 39: 319–342. doi:10.1080/0144929X.2019.1620334.

Kaushal, R., S. Chandok, P. Jain, P. Dewan, N. Gupta, and P. Kumaraguru. 2017. "Nudging Nemo: Helping Users Control Linkability Across Social Networks." In: *International Conference on Social Informatics*. 477–490, Springer. doi:https://doi.org/10.1007/978-3-319-67256-4_38.

Kim, B., D. Y. Lee, and B. Kim. 2019. "Deterrent Effects of Punishment and Training on Insider Security Threats: A Field Experiment on Phishing Attacks." *Behaviour and Information Technology*. doi:10.1080/0144929X.2019.1653992.

Knijnenburg, B. 2017. "Privacy? I Can't Even! Making a Case for User-Tailored Privacy." *IEEE Security and Privacy* 15: 62–67. doi:10.1109/MSP.2017.3151331.

Kroll, T., and S. Stieglitz. 2019. "Digital Nudging and Privacy: Improving Decisions About Self-Disclosure in Social Networks." *Behaviour and Information Technology* doi:10.1080/0144929X.2019.1584644.

Kwon, B. C., and B. Lee. 2016. "A Comparative Evaluation on Online Learning Approaches using Parallel Coordinate Visualization." *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 993–997. doi:10.1145/2858036.2858101.

Li, L., E. Berki, M. Helenius, and S. Ovaska. 2014. "Towards a Contingency Approach with Whitelist- and Blacklist-based Anti-phishing Applications: What Do Usability Tests Indicate?." *Behaviour and Information Technology* 33: 1136–1147. doi:10.1080/0144929X.2013.875221.

McKenna, S., D. Staheli, and M. Meyer. 2015. "Unlocking User-Centered Design Methods for Building Cyber Security Visualizations." *2015 IEEE Symposium on Visualization for Cyber Security (VizSec)*, 1–8. doi:10.1109/VIZSEC.2015.7312771.

Merkel, C., and R. Wiczorek. 2012. "Does Higher Security Always Result in Better Protection? An Approach for Mitigating the Trade-off Between Usability and Security. Human Factors: A View From an Integrative Perspective." *Proceedings HFES Europe Chapter Conference Toulouse*, 29–42.

Micallef, N., M. Just, L. Baillie, and M. Alharby. 2017. "Stop Annoying Me! An Empirical Investigation of the App Privacy Notifications." *Proceedings of the 29th Australian Conference on Computer–Human-Interaction*, 371–375. ACM. doi:10.1145/3152771.3156139.

Mohamed, M. A., J. Chakraborty, and J. Dehlinger. 2017. "Trading Off Usability and Security in User Interface Design Through Mental Models." *Behaviour and Information Technology* 36: 493–516. doi:10.1080/0144929X.2016.1262897.

Pacini, R., and S. Epstein. 1999. "The Relation of Rational and Experiential Information Processing Styles to Personality, Basic Beliefs, and the Ratio-bias Phenomenon." *Journal of Personality and Social Psychology* 76: 972–987. doi:10.1037/0022-3514.76.6.972.

Peer, E., S. Egelman, M. Harbach, N. Malkin, A. Mathur, and A. Frik. 2019. "Nudge Me Right: Personalizing Online Nudges to People's Decision-Making Styles." *Computers in Human Behavior* 109: 106347.

Ramesh, G., K. Selvakumar, and A. Venugopal. 2017. "Intelligent Explanation Generation System for Phishing Webpages by Employing An Inference System." *Behaviour and Information Technology* 36: 1244–1260. doi:10.1080/0144929X.2017.1369569.

Reeder, R., A. P. Felt, S. Consolvo, N. Malkin, C. Thompson, and S. Egelman. 2018. "An Experience Sampling Study of User Reactions to Browser Warnings in the Field." *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM* 1: 1–13. doi:10.1145/3173574.3174086.

Renaud, K., and V. Zimmerman. 2017. "Enriched Nudges Lead to Stronger Password Replacements…but Implement Mindfully." *2017 Information Security for South Africa – Proceedings of the 2017 ISSA Conference*, 1–9. doi:10.1109/ISSA.2017.8251779.

Renaud, K., V. Zimmerman, J. Maguire, and S. Draper. 2017. "Lessons Learned From Evaluating Eight Password Nudges in the Wild." In: *The LASER Workshop: Learning from Authoritative Security Experiment Results (LASER 2017)*. 25–37, USENIX Association. https://www.usenix.org/conference/laser2017/presentation/renaud.

Renaud, K., and V. Zimmermann. 2018a. "Ethical Guidelines for Nudging in Information Security & Privacy." *International Journal of Human Computer Studies* 120: 22–35. doi:10.1016/j.ijhcs.2018.05.011.

Renaud, K., and V. Zimmermann. 2018b. "Guidelines for Ethical Nudging in Password Authentication." *SAIEE Africa Research Journal* 109: 102–118.

Saad, T., and F. Khan. 2016. "Nudging Pakistani Users Towards Privacy on Social Networks." 2016 SAI Computing Conference (SAI), 1147–1154. doi:10.1109/SAI.2016.7556122.

Santos, J. R., Y. Y. Haimes, and C. Lian. 2007. "A Framework for Linking Cybersecurity Metrics to the Modeling of Macroeconomic Interdependencies." *Risk Analysis: An International Journal* 27: 1283–1297. doi:10.1111/j.1539-6924.2007.00957.x.

Savola, R. M., and P. Heinonen. 2011. "A Visualization and Modeling tool for Security Metrics and Measurements Management." *2011 Information Security for South Africa -- Proceedings of the ISSA 2011 Conference*. doi:10.1109/ISSA.2011.6027518.

Scott, S., and R. Bruce. 1995. "Decision-making Style: The Development and Assessment of a New Measure." *Educational and Psychological Measurement* 55: 818–831.

Segreti, S. M., W. Melicher, S. Komanduri, D. Melicher, R. Shay, B. Ur, L. Bauer, N. Christin, L. F. Cranor, and M. L. Mazurek. 2017. "Diversify to Survive: Making Passwords Stronger With Adaptive Policies." In: *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. 1–12, Santa Clara, CA: USENIX Association. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/segreti.

Story, P., D. Smullen, A. Acquisti, L. F. Cranor, N. Sadeh, and F. Schaub. 2020. "From Intent to Action: Nudging Users Towards Secure Mobile Payments." In: *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, 379–415.

Tan, J., L. Bauer, N. Christin, and L. F. Cranor. 2020. "Practical Recommendations for Stronger, More Usable Passwords Combining Minimum-Strength, Minimum-Length, and Blocklist Requirements." In: *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 1407–1426. New York, NY: Association for Computing Machinery. doi:10.1145/3372297.3417882.

Thaler, R., and C. Sunstein. 2009. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven, CT: Penguin.

Turland, J., L. Coventry, D. Jeske, P. Briggs, and A. van Moorsel. 2015. "Nudging Towards security: Developing an Application for Wireless Network Selection for Android Phones." *Proceedings of the 2015 British HCI Conference on -- British HCI '15*, 193–201. doi:10.1145/2783446.2783588.

Tussyadiah, I., S. Li, and G. Miller. 2019. "Privacy Protection in Tourism: Where we are and Where we Should be Heading for." *Information and Communication Technologies in Tourism*, 278–290. https://doi.org/10.1007/978-3-030-05940-8_22.

Ur, B., F. Alfieri, M. Aung, L. Bauer, N. Christin, J. Colnago, L. F. Cranor, et al. 2017. "Design and Evaluation of a Data-Driven Password Meter." *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3775–3786. ACM. doi:10.1145/3025453.3026050.

Ur, B., J. Bees, S. M. Segreti, L. Bauer, N. Christin, and L. F. Cranor. 2016. "Do Users' Perceptions of Password Security Match Reality?" In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3748–3760. ACM. doi:10.1145/2858036.2858546.

Ur, B., P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. L. Mazurek, T. Passaro, et al. 2012. "How Does Your Password Measure up? The Effect of Strength Meters on Password Creation." in: *21st USENIX Security Symposium (USENIX Security 12)*, 65–80. Bellevue, WA: USENIX Association. https://www.usenix.org/conference/usenixsecurity12/technical-sessions/presentation/ur.

Ur, B., F. Noma, J. Bees, S. M. Segreti, R. Shay, L. Bauer, N. Christin, and L. F. Cranor. 2015. ""i added '!' at the end to Make it Secure": Observing Password Creation in the Lab." In: *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, 123–140. Ottawa: USENIX Association. https://www.usenix.org/conference/soups2015/proceedings/presentation/ur.

Van Bavel, R., N. Rodríguez-Priego, J. Vila, and P. Briggs. 2019. "Using Protection Motivation Theory in the Design of Nudges to Improve Online Security Behavior." *International Journal of Human Computer Studies* 123: 29–39. doi:10.1016/j.ijhcs.2018.11.003.

Vance, A. 2010. "If Your Password Is 123456, Just Make It HackMe." https://www.nytimes.com/2010/01/21/technology/21password.html.

Van Someren, M. W., Y. F. Barnard, and J. A. C. Sandberg. 1994. *The Think Aloud Method: a Practical Approach to Modelling Cognitive Processes*. London: Academic Press.

Veras, R., C. Collins, and J. Thorpe. 2014. "On the Semantic Patterns of Passwords and their Security Impact." *NDSS*, 23–26. doi:10.14722/ndss.2014.23103.

Veras, R., J. Thorpe, and C. Collins. 2012. "Visualizing Semantics in Passwords: The Role of Dates." *Proceedings of the Ninth International Symposium on Visualization for Cyber Security -- VizSec '12*, 88–95. doi:10.1145/2379690.2379702.

Wachner, J., M. Adriaanse, and D. De Ridder. 2020. "The Influence of Nudge Transparency on the Experience of Autonomy." *Comprehensive Results in Social Psychology,* 3: 1–15.

Wang, Y., P. G. Leon, A. Acquisti, L. F. Cranor, A. Forget, and N. Sadeh. 2014. "A Field Trial of Privacy Nudges for Facebook." *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, 2367–2376. doi:10.1145/2556288.2557413.

Wang, D., Q. Yang, A. Abdul, and B. Y. Lim. 2019. "Designing Theory-Driven User-Centric Explainable AI." *2019 CHI Conference on Human Factors in Computing Systems Proceedings*, 1–15. doi:10.1145/3290605.3300831.

Ware, C. 2012. *Information Visualization: Perception for Design*. Waltham, MA: Elsevier.

Wheeler, D. 2016. "zxcvbn: Low-Budget Password Strength Estimation zxcvbn." *25th USENIX Security Symposium*, 157–173. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/wheeler.

Wisniewski, P. J., B. P. Knijnenburg, and H. R. Lipford. 2017. "Making Privacy Personal: Profiling Social Network Users to Inform Privacy Education and Nudging." *International Journal of Human Computer Studies* 98: 95–108. doi:10.1016/j.ijhcs.2016.09.006.

Yan, J., B. Alan, R. Anderson, and A. Grant. 2004. "Password Memorability and Security: Empirical Results." *IEEE Security and Privacy* 2: 25–31. doi:10.1109/MSP.2004.81.

Yu, X., and Q. Liao. 2016. "User Password Repetitive Patterns Analysis and Visualization." *Information and Computer Security* 24: 93–115. doi:10.1108/ICS-06-2015-0026.

Zhang, B., and H. Xu. 2016. "Privacy Nudges for Mobile Applications: Effects on the Creepiness Emotion and Privacy Attitudes." *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* 27: 1676–1690. doi:10.1145/2818048.2820073.

Zimmerman, S., A. Thorpe, C. Fox, and U. Kruschwitz. 2019. "Privacy Nudging in Search: Investigating Potential Impacts." *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval. ACM* 3: 283–287. doi:10.1145/3295750.3298952.

Zimmermann, V., and N. 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: 26–44. doi:10.1016/j.ijhcs.2019.08.006.

Appendices

Appendix 1. Survey instrument

A.1. Demographic items

- What gender do you identify with?
  - female
  - male
  - other
- – What is your age (in years)?
  - younger than 18
  - 18–29
  - 30–39
  - 40–49
  - 50–59
  - 60 or older
- In which federal state do you currently live? [choose from list of all German federal states]
- Please indicate your highest level of education
  without a school diploma / certificate of secondary education
  - general certificate of secondary education
  - qualification for university entrance / university degree
- What is your monthly household income?
  - under *e*2000
  - *e*2000 to 4000
  - above *e*4000
- Do you like to play role-playing games? [yes/no/no answer]

A.2. Psychometric tests

- General Decision Making Style (subscales R = *rational* and D = *dependent*) (Scott and Bruce 1995) [scoring: 1 (strongly disagree) to 5 (strongly agree)]
  - I rarely make important decisions without consulting other people. (D)
  - I double-check my information sources to be sure I have the right facts before making decisions. (R)
  - I use the advice of other people in making my important decisions. (D)
  - I make decisions in a logical and systematic way. (R)
  - I like to have someone to steer me in the right direction when I am faced with important decisions. (D)
  - My decision making requires careful thought. (R)
  - When making a decision, I consider various options in terms of a specific goal. (R)
  - I often need the assistance of other people when making important decision. (D)
  - If I have the support of others, it is easier for me to make important decisions. (D)
  - I make decisions in a logical and systematic way. (R)
- Rational-Experiential Inventory (subscales RA = *rational ability* and RE = *rational engagement* (Pacini and Epstein 1999) [scoring: 1 (strongly disagree) to 5 (strongly agree)])
  - I try to avoid situations that require thinking in a depth about something. (RE)
  - I'm not that good at figuring out complicated problems. (RA)
  - I enjoy intellectual challenges. (RE)
  - I am not very good at solving problems that require careful logical analysis. (RA)
  - I don't like to have to do a lot of thinking. (RE)
  - I enjoy solving problems that require hard thinking. (RE)
  - Thinking is not my idea of an enjoyable activity. (RE)
  - I am not a very analytical thinker. (RA)
  - Reasoning things out carefully is not one of my strong points. (RA)
  - I prefer complex problems to simple problems. (RE)
  - Thinking hard and for a long time about something gives me little satisfaction. (RE)
  - I don't reason well under pressure. (RA)
  - I am much better at figuring things out logically than most people. (RA)
  - I have a logical mind. (RA)
  - I enjoy thinking in abstract terms. (RE)
  - I have no problem thinking things through carefully. (RA)
  - Using logic usually works well for me in figuring out problems in my life. (RA)
  - Knowing the answer without having to understand the reasoning behind it is good enough for me. (RE)
  - I usually have clear, explainable reasons for my decisions. (RA)
  - Learning new ways to think would be very appealing to me. (RE)

A.3. Adapted system usability scale

Based on the System Usability Scale (SUS) by Brooke (1996). [scoring: 1 (strongly disagree) to 5 (strongly agree)]

- I think that I would like to use this visualization frequently.
- I found the visualization unnecessarily complex.
- I thought the visualization was easy to use.
- I think that I would need the support of a technical person to be able to use this visualization.
- By means of the visualization I know what to do to improve the strength of my password. [replaces item of original SUS]
- I thought there was too much inconsistency in this visualization.
- I would imagine that most people would learn to use this visualization very quickly.
- I found the visualization very cumbersome to use.
- I felt very confident using the visualization.
- I needed to learn a lot of things before I could get going with this visualization.
- By means of the visualization I can evaluate the strength of my password. [added to original SUS]

A.4. Items on attitudes

A nudge is an instrument to alter people's behaviour. The individual is gently steered in a specific direction without forbidding any alternative options. In the context of health, an exemplary nudge can be arranging fruits and vegetables in the school cafeteria at eye level. Thus the students are animated to choose healthier food.

Also, in the context of cybersecurity nudges can be applied to steer people's behaviour in a more secure direction. Nudges can, for instance, remind of backups of important data, warn against phishing mails or indicate when a password is not

strong enough. That can, for example, take place using images, slogans or colours among other.

Please indicate to what extent you agree with the following statements. There is no right or wrong, we are only interested in your opinion.

- It is important to me to understand how the assessment of the strength of my password was calculated online. [scoring: 1 (strongly disagree) to 5 (strongly agree)]
- It is important to me to realise when someone tries to nudge me online. [scoring: 1 (strongly disagree) to 5 (strongly agree)]
- Nudging is useful for the following contexts: [scoring: 1 (strongly disagree) to 5 (strongly agree), no answer]
  - password creation for important accounts
  - password creation for unimportant accounts
- Do you consider nudges to be a sensible way to steer online behaviour in a secure direction or do you see any risks? [free-text format]

Appendix 2. Additional statistical results

Table A1. Significant MANOVA results regarding nudging efficacy (without interaction effects).

Dependent variable	Independent variable	Df	*F* value	*p* value
length	nudge type	2	16.07	<.0001
length	REI_engagement score	1	4.34	.0375
length	GDMS_dependent score	1	6.58	.0105
length	GDMS_rational score	1	9.14	.0026
digit	nudge type	2	4.37	.0129
digit	REI_engagement score	1	5.62	.0180
digit	gaming	2	3.28	.0382
lower case	nudge type	2	2.59	.0758
lower case	GDMS_dependent score	1	3.88	.0491
upper case	nudge type	2		.0129
upper case	REI_engagement score	1	5.62	.0180
upper case	gaming	2	3.28	.0382
special character	nudge type	2	10.84	<.0001
special character	REI_engagement score	1	10.28	.0014
special character	GDMS_dependent score	1	5.51	.0191
special character	GDMS_rational score	1	6.498	.0110
spec. character/digit in middle	nudge type	2	11.39	<.0001
spec. character/digit in middle	REI_ability score	1	3.35	.0675
spec. character/digit in middle	REI_engagement score	1	14.01	.0002
spec. character/digit in middle	GDMS_rational score	1	3.78	.0521
zxcvbn-score	nudge type	2	18.19	<.0001
zxcvbn-score	REI_engagement score	1	8.29	.0040
zxcvbn-score	GDMS_rational score	1	12.31	.0005

Table A2. Results of post-hoc Tukey's HSD tests regarding nudging efficacy Part I.

Dependent variable	Indep. variable	Combination	Diff of means	Adj. *p* value
length	nudge type	None-Meter	−1.66	<.0001
length	nudge type	Radar-Meter	0.28	.7381
length	nudge type	Radar-None	1.92	<.0001
length	REI_eng.	low-high	−0.75	.0410
length	GDMS_dep.	low-high	0.77	.0118
length	GDMS_rat.	low-high	−1.83	.0031
digit	nudge type	None-Meter	−0.45	.0286
digit	nudge type	Radar-Meter	−0.01	.9996
digit	nudge type	Radar-None	0.44	.0292
digit	REI_eng.	low-high	−0.40	.0202
digit	gaming	No-Yes	−0.34	.0456
digit	gaming	No answer-Yes	−0.61	.4060
digit	gaming	No answer-No	−0.45	.8429
lower case	nudge type	None-Meter	−0.56	.2675
lower case	nudge type	Radar-Meter	0.23	.7972
lower case	nudge type	Radar-None	0.79	.0696
lower case	GDMS_dep.	low-high	0.57	.0531
upper case	nudge type	None-Meter	−0.35	.0145
upper case	nudge type	Radar-Meter	−0.07	.8232
upper case	nudge type	Radar-None	0.28	.0680
upper case	REI_ab.	low-high	−0.33	.0492
upper case	REI_eng.	low-high	−0.26	.0412
upper case	GDMS_rat.	low-high	−0.46	.0267
upper case	gaming	No-Yes	0.24	.0536
upper case	gaming	No answer-Yes	−0.28	.6978
upper case	gaming	No answer-No	−0.52	.2828

Table A3. Results of post-hoc Tukey's HSD tests regarding nudging efficacy part II.

Dependent variable	Indep. variable	Combination	Diff of means	Adj. *p* value
special character	nudge type	None-Meter	−0.255	.0111
special character	nudge type	Radar-Meter	0.15	.2155
special character	nudge type	Radar-None	0.40	<.0001
special character	REI_eng.	low-high	−0.255	.0111
special character	GDMS_dep.	low-high	0.17	.0213
special character	GDMS_rat.	low-high	−0.37	.0125
spec. char./digit in middle	nudge type	None-Meter	−0.28	.0015
spec. char./digit in middle	nudge type	Radar-Meter	0.08	.5626
spec. char./digit in middle	nudge type	Radar-None	0.37	<.0001
spec. char./digit in middle	REI_ab.	low-high	−0.19	.0678
spec. char./digit in middle	REI_eng.	low-high	−0.30	.0002
spec. char./digit in middle	GDMS_rat.	low-high	−0.26	.0565
zxcvbn-score	nudge type	None-Meter	−0.46	<.0001
zxcvbn-score	nudge type	Radar-Meter	0.02	.9795
zxcvbn-score	nudge type	Radar-None	0.48	<.0001
zxcvbn-score	REI_eng.	low-high	−0.26	.0048
zxcvbn-score	GDMS_rat.	low-high	−0.52	.0006