GazeMeter: Exploring the Usage of Gaze Behaviour to Enhance Password Assessments

YASMEEN ABDRABOU, Bundeswehr University Munich, Germany
AHMED SHAMS, German University in Cairo, Egypt
MOHAMED MANTAWEY, German University in Cairo, Egypt
ANAM AHMAD KHAN, University of Melbourne, Australia
MOHAMED KHAMIS, University of Glasgow, United Kingdom
FLORIAN ALT, Bundeswehr University Munich, Germany
YOMNA ABDELRAHMAN, Bundeswehr University Munich, Germany

We investigate the use of gaze behaviour as a means to assess password strength as perceived by users. We contribute to the effort of making users choose passwords that are robust against guessing-attacks. Our particular idea is to consider also the users' understanding of password strength in security mechanisms. We demonstrate how eye tracking can enable this: by analysing people's gaze behaviour during password creation, its strength can be determined. To demonstrate the feasibility of this approach, we present a proof of concept study ($N = 15$) in which we asked participants to create weak and strong passwords. Our findings reveal that it is possible to estimate password strength from gaze behaviour with an accuracy of 86% using Machine Learning. Thus, we enable research on novel interfaces that consider users' understanding with the ultimate goal of making users choose stronger passwords.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; **Interactive systems and tools**.

Additional Key Words and Phrases: Eye-tracking, Gaze Behaviour, Password Meters, Password Strength

ACM Reference Format:
Yasmeen Abdrabou, Ahmed Shams, Mohamed Mantawey, Anam Ahmad Khan, Mohamed Khamis, Florian Alt, and Yomna Abdelrahman. 2021. GazeMeter: Exploring the Usage of Gaze Behaviour to Enhance Password Assessments. In *2021 Symposium on Eye Tracking Research and Applications (ETRA '21 Full Papers), May 25–27, 2021, Virtual Event, Germany*. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3448017.3457384

1 INTRODUCTION

Text-based passwords are still among the most commonly used means for authentication. The most important reasons for this are that users are familiar with this approach – hence, requiring only little learning – and that such schemes are easy to implement. At the same time, a well-known issue is that, despite the many existing approaches and tools to support the use of stronger passwords, people are still not selecting strong passwords [Florêncio et al. 2014].

This is, on one hand, a result of usability issues [Bonneau et al. 2015; Egelman et al. 2013; Florêncio et al. 2014]. On the other hand, it has been understood by the usable security community that many users have a wrong perception of factors contributing to password strength [Stobert and Biddle 2016; Ur et al. 2016, 2015]. This is mostly due to password creation rules and policies being inconsistent and misleading, ultimately resulting in such wrong perceptions of password strength [Das et al. 2014; Leonhard and Venkatakrishnan 2007; Seitz et al. 2017; Wang and Wang 2015].

As one solution to this, many services (in particular, web sites) employ password meters, providing people an estimate of their chosen password's strength. This is done through visual aids that provide instant feedback on password strength in the form of coloured bars. Instead of forcing users to choose stronger passwords, password meters nudges users to rethink their password choice. However, similar to policies, also password meters often suffer from inconsistent ratings across different meters. For example, one meter might consider "password$1" a strong choice whereas another meter might rate the security of this password much lower [de Carné de Carnavalet and Mannan 2014]. In any case, a system employing a password meter might still accept a password, despite being relatively insecure. This is likely to create a wrong perception among users that also the use of weaker passwords is acceptable, regardless of existing concerns and will confuse them, ultimately reducing credibility and understandability of password meters.

This demonstrates the need to develop better mechanisms to make users choose stronger passwords. We see particular potential in mechanisms, that not only take into account the actual strength of the password, but also how users perceive it. This creates two important prerequisites: firstly, a system needs to be able to infer the perceived strength of the users' chosen passwords; secondly, the system should be capable of doing so without creating additional effort for the user. As a result, novel approaches to increase password strength could be created or existing ones, such as password meters, be enhanced. We outline ideas after the contribution statement.

We address this by showing that password strength can be inferred implicitly, that is, without any need for interaction by the user, from gaze behaviour upon password creation. Our approach is based on the assumption that the cognitive processes while coming up with weak or strong passwords differently influence users' physiological response and behaviour, reflected, for example, in their gaze behaviour. We implicitly monitored and analysed users' gaze behaviour while creating weak and strong passwords in a lab study. We investigated two input devices – laptops and smartphones – to compare differences in gaze behaviour among the most frequently used input devices. We report on the performance and compare different machine learning classifiers. In particular, we investigated, user-dependent and user-independent classifiers. Besides, we validate the collected passwords by comparing their entropy against the zxcvbn password meter [Wheeler 2016] which is a password strength estimator using pattern matching and conservative estimation.

Our findings demonstrate the potential of eye tracking for unobtrusive classification of password strength. We found a promising accuracy of up to 86% for personalised classifiers on smartphones and 80% on laptops. We found that the average pupil diameter, average saccadic duration, fixation duration, and the duration spent while entering the passwords are good features for detecting passwords strength from users' gaze behaviour.

Contribution Statement We propose a novel approach for classifying users' password strength (weak vs. strong) by monitoring users' gaze. Secondly, we present a proof-of-concept implementation and evaluate it in a user study (N=15).

We believe the research community and practitioners can benefit from our work in several ways. Our approach supports the design of novel interfaces that make people use stronger passwords. Such designs can either nudge the user towards the use of more secure passwords (similar to traditional password meters), with the difference that the approach is not based on password entropy. Or novel designs can use knowledge on password strength for entirely new concepts that, for example, only allow users to register passwords if their strength match the sensitivity of the data that need to be protected. We see two particular strengths of our approach: firstly, it is *independent of the underlying authentication mechanism*. Hence, our approach does not require any knowledge about the actual password (as opposed to traditional password meters), hence minimising the attack surface; secondly, concepts can be implemented *independent of the input device*. For example, using a mobile eye tracker, the strength of a password entered on a desktop computer can be

assessed and recommendations for a better password be provided on a smartwatch. Finally, future work could apply our concept to other knowledge-based authentication schemes, such as lock patterns or image-based approaches.

2 RELATED WORK

Our work draws from several strands of prior research, most importantly research on password strength meters and gaze behaviour in the context of passwords.

2.1 Password Strength Meters

The use of password strength meters was adopted a decade ago [Weir et al. 2010]. Many studies investigated the effectiveness of using password meters on the security and memorability of passwords. Work by Ur et al. [de Carné de Carnavalet and Mannan 2014] showed that participants believe that adding an exclamation mark at the end of their passwords would make it stronger. Participants also believed that having keyboard patterns or adding their pet name in the password is an asset for a strong password.

In 2013, Egelman et al. [Egelman et al. 2013] examined whether the use of password meters influenced users' password strength or not. The authors asked participants to first change their real passwords according to the presence of the password meters, next to change an important account password, and finally to change an unimportant account password. They found that password meters significantly enhanced users' generated passwords for their real accounts and important accounts. However, for non-important accounts, password meters did not have an affect. The authors concluded that the use of password meters is only effective if the user is forced to change or create a password for an important account.

Further research by Ur et al. showed that also the appearance of the password meter affects the choice of passwords [Ur et al. 2012]. For example, meters without visual bars gave participants the impression that it is not important to enter a strong password and, hence, caused participants to put less effort in satisfying the meter's requirement. In contrast, participants who saw more lenient meters tried to fulfil the meter requirements and were reluctant to choosing passwords a meter deemed as "bad" or "poor".

In 2014, Shay et al. [Shay et al. 2014], studied the effect of password length on password strength. They found that policies requiring longer passwords reduces the percentage of easy-to-guess passwords. They also found that enforcing combinations of certain requirements and increased password length led to stronger passwords and was more usable compared to traditionally complex policies. Later, Shay et al. [Shay et al. 2015], studied the usability of feedback and guidance mechanisms for password meters. They found that service providers should present password requirements in combination with feedback to increase usability. However, feedback needs to be designed carefully, as the same requirements can have different security and usability effects depending on the way they are presented.

In 2017, Ur et al. [Ur et al. 2017] proposed a 'Data-Driven Password Meters'. The meter communicates up to 3 ways to the user how the entered password can be enhanced. The results showed that data-driven meters with detailed feedback led users to create more secure, yet equally memorable passwords, compared to normal meters with a strength bar indicator. Research by Dupuis et al. [Dupuis and Khan 2018], studied the effect of changing the feedback on generated passwords' strength. Instead of indicating the actual password strength, they provided a comparison of the strength to passwords of other users. For example, instead of showing *weak password*, they showed *weak compared to other users*. The authors report that by changing the feedback mechanism and comparing users' passwords to others, people generated stronger passwords.

2.2 Gaze Behaviour and Passwords

Eye trackers are becoming ubiquitous. Today, they are already integrated in some laptops[1] or embedded as front-facing depth cameras in some smartphones[2]. Future generations of laptops and smartphones may ship with built-in eye tracker as default feature. These may benefit from decades of research that investigated the use of eye gaze as an interaction modality [Forget et al. 2010; Kumar et al. 2007; Majaranta and Räihä 2007], hybrid modality [Abdrabou et al. 2019; Khamis et al. 2016, 2017] and as a behavioural modality. Gaze behaviour has been integrated in many areas, including but not limited to detecting personality traits [Hoppe et al. 2018], detecting activity recognition [Bulling et al. 2011] and measuring cognitive load [Henderson et al. 2013].

In particular, security mechanisms might benefit from eye gaze [Katsini et al. 2020]. Eye gaze has been used for continuous verification [Abdulin and Komogortsev 2015; Cantoni et al. 2018; Zhang et al. 2018] and implicit identification [Bayat and Pomplun 2018; Cantoni et al. 2015; Vitonis and Hansen 2014]. In 2018, Katsini et al. [Katsini et al. 2018a], investigated users' visual behaviour and how it affects the strength of the created picture passwords. They used cognitive style theories to interpret their results. They found that users with different cognitive styles followed different patterns of visual behaviour, which affected the strength of the created passwords. Furthermore, The authors introduced and studied adaptive characteristics of authentication mechanism, aiming to assist user groups following different cognitive styles to create more secure passwords. The results confirmed that adaptive mechanisms based on different cognitive and visual behaviour enables new ways of improving password strength in graphical user authentication.

Other work by Katsini et al. [Katsini et al. 2018b], studied the feasibility of estimating the strength of user-created graphical passwords based on gaze behaviour during password composition. The authors used unique fixations on each area of interest (AOI) and the total fixation duration per AOI. The authors also investigate whether gaze-based entropy is a credible predictor of password strength. Their results revealed a strong positive correlation between password strength and gaze-based entropy. This suggests that the proposed gaze-based metric enables the strength of the password to be predicted in an unobtrusive manner and, thus, help users create stronger passwords. We adopted a similar strategy for detecting password strength from users' gaze. In contrast to prior work we focus on text-based passwords (instead of graphical ones) and we assess password strength as perceived by users (as opposed to password strength as assessed by a system).

As discussed, throughout the years, password meters and heuristics have biased users' choice of passwords and forced them to adopt similar strategies for passwords creation. This yields a major security risk as most of the users creates similar passwords which makes them more vulnerable to attacks. With the ubiquity of eye trackers and by proving that eye gaze behaviour can act as a picture password strength meter, we propose adopting the same idea of using eye gaze behaviour to estimate text-based passwords' strength. We hypothesise that users behaviour (reflected in the gaze data) while creating a strong password is different than while creating weak passwords and it can be used as a new behavioural aspect.

3 EYE TRACKING FOR PASSWORD STRENGTH CLASSIFICATION

Previous work showed that security mechanisms can generally strongly benefit from the use of eye gaze data. As previously mentioned, this becomes possible through eye trackers being increasingly available in situations in which security-related tasks (such as authentication) is performed. Hereby, a particular strength of eye tracking is that

[1] https://gaming.tobii.com/products/laptops/
[2] https://www.apple.com/newsroom/2017/09/the-future-is-here-iphone-x/

assistance during security-critical tasks can be provided in an unobtrusive, implicit manner, i.e. a system can make use of gaze data without the need for action from the user.

In this work, we investigate a novel application area of using gaze data in security-critical contexts, that is the implicit assessment of password strength as perceived by users. In particular, we focus on the distinction between weak and strong passwords with the ultimate goal of supporting the design of future mechanisms that use this knowledge for interventions that make users chose stronger passwords.

3.1 Password Strength

Password strength can be assessed in different ways. Traditionally the theoretical password space was used to determine password strength, that is the overall number of possible passwords. However, it is today well understood that passwords are not uniformly distributed over the password space, since certain passwords are more likely to be chosen by users than others (for example, 'password' or '123456'. Hence, researchers today rather consider the practical password space, that is the number of actually used passwords. This password space is generally assessed through empirical studies.

Password strength estimators, such as *zxcvbn* are considering this fact. Hereby, strength is determined through the average number of guesses required to identify a password (a so-called guessing attack). The mentioned password estimator, which today serves as a standard way of estimating password strength in security research, classifies passwords into 5 categories: (1) too guessable passwords can be identified through less than 10^3 guesses. (2) very guessable passwords, which protect from throttled online attacks, require about 10^6 guesses. (3) somewhat guessable passwords prevent unthrottled online attacks, requiring on average 10^8 guesses. (4) Safely unguessable passwords provide moderate protection from offline slow-hash attack scenarios (10^{10} guesses). (5) Finally, very unguessable passwords provide strong protection by requiring more than 10^{10} guesses.

In the context of our work we consider weak passwords any password that requires on average below 10^7 guesses, according to zxcvbn. Strong passwords are such that require on average more than 10^7 guesses.

3.2 Perceived Password Strength

As laid out in the motivation of our work, a major challenge in usable security research is the mismatch between the password strength as determined by a strength estimator (we refer to this as the *actual password strength* and the strength as perceived by users *perceived password strength*). Figure 1 demonstrates this mismatch and its implications.

Optimally, the way users perceive the strength of their passwords would match the actual password strength (i.e., both strong – upper left, both weak – lower right). This would allow them to make a reasonable decision, whether or not their password is appropriate for the type of data they seek to protect. What is now interesting are cases in which actual and perceived passwords strength do not match. In the case where the actual password is strong, but perceived weak by users, no harm would be caused, but it might be worthwhile to explain users their misconception. More problematic is the other case in which the password is perceived as strong by users but is actually weak. In this case it might be useful to both explain this misconception (and the reasons for it) to the user but additionally to also support or even require them to create stronger passwords.

Our work is meant to particularly identify cases where actual password strength and perceived password strength are at odds. In this way, we enable researchers to come up with interventions that address the respective cases.

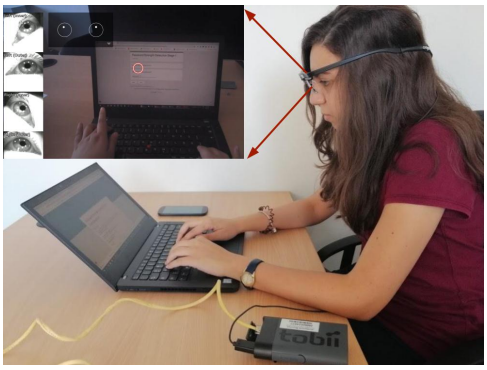

Fig. 1. Experiment study setup consisting of laptop, wearable eye tracker and the smartphone used. Top Left: gaze monitoring while creating passwords viewed from Tobii pro glasses controller.

Table 1. Differences between actual password strength and perceived password strength and potential use cases.

Actual Strength/ Perceived Strength	**Strong**	**Weak**
Strong	No action is required.	Need to clarify misconception / motivate users to chose stronger password. .
Weak	Opportunity to explain misconception to users	Opportunity to make users consider whether password strength is appropriate.

4 STUDY

To demonstrate that it is possible to infer perceived password strength from gaze data, we conducted a proof of concept user study. We recorded participants' eye gaze data while creating weak and strong passwords on two input modalities: laptops and touchscreen smartphones. We chose to include different input devices to understand the the influence on gaze movements, in particular, or eye movements between keyboard and screen.

4.1 Design

We applied a repeated-measures design, where all participants experienced all conditions. Participants were asked to enter 24 passwords (6 weak and 6 strong) on both laptop and smartphone.The order of the devices and the password they should create were counterbalanced using a Latin Square. Participants were advised to neither reuse passwords they were already using beforehand nor to reuse passwords they came up with for the study.

4.2 Apparatus

The experimental setup consisted of Lenovo T480[3] and Yotaphone[4] as input devices (Figure 1). For the eye tracker we used the Tobii Pro Glasses[5], connected to a Lenovo T440s[6] using with the Tobii glasses controller[7]. We decided on a wearable eye tracker in order to use the same hardware across all conditions. Also, this allowed us to assess participants' pupil diameter. Deployed systems may rely integrated cameras such as front facing depth cameras in smartphones [Khamis et al. 2018]. We implemented a simple web page interface showing the task and login interface.

[3]Lenovo T480: https://www.lenovo.com/us/en/laptops/thinkpad/thinkpad-t-series/ThinkPad-T480/p/22TP2TT4800
[4]Yotaphone: https://www.cect-shop.com/de/yota-yotaphone-3-plus.html
[5]Tobii Pro Glasseshttps://www.tobiipro.com/product-listing/tobii-pro-glasses-2/
[6]Lenovo T440s: https://www.lenovo.com/gb/en/laptops/thinkpad/t-series/t440s/
[7]Tobii Glasses Controller: https://www.tobiipro.com/learn-and-support/learn/steps-in-an-eye-tracking-study/setup/installing-tobii-glasses-controller/

4.3 Recruiting and Procedure

We recruited 15 participants (5 males) via University mailing lists. The age varied from 22 to 31 (Mean = 24.27; SD = 2.91). Participants had different backgrounds (CS, engineering, landscape design) and different nationalities (Spain, China, Bangladesh, Pakistan, Egypt, Germany). They had basic to average eye-tracking experience. Nobody wore glasses.

After arriving at the lab, participants signed a consent form. Then we explained the purpose of the study. After that, we calibrated the eye tracker using Tobii's one-point calibration. We then asked them to begin creating passwords. After each password, we asked the participants to rate the password's strength on a Likert-scale (1=very weak; 5=very strong). After creating password on both devices, we interviewed them about what they though characterizes a strong password. The study lasted approximately 20 minutes. Participants were compensated with 5 EUR.

4.4 Limitations

In our study, people did not create passwords to protect real data. Yet, prior research showed that people in such studies still create realistic passwords. We specifically focused on cases where people created new passwords. In reality, password reuse is a common strategy to cope with the issue of having to memorize too many passwords. The effect of this strategy on perceived password strength estimation could be subject to future work. We acknowledge that the sample for our proof-of-concept study was rather small. At the same time, it is in line with prior studies including password creation tasks [Forget et al. 2008; Notoatmodjo and Thomborson 2009; Rinn et al. 2015]. Also, asking people to create multiple passwords still allowed us to collect a data set appropriate for the employed machine learning techniques (cf. the confusion matrix in Figure 4).

5 METHODOLOGY

In this section, we describe the step-by-step process to derive perceived password strength from eye gaze.

5.1 Statistical Analysis and Password Strength Estimation

To validate the collected passwords, we analysed and compared passwords entropy and user rated password strength against the zxcvbn password strength estimator [Wheeler 2016] (details can be found in Section 6.1.1 and 6.1.2). We normalised the zxcvbn password strength estimator score to the range of 1 to 5 and used it to classify passwords into weak and strong. We used a cut-off score of 2.5 for differentiating between weak and strong passwords, i.e. passwords with a score of 1 to 2.5 are considered weak, whereas passwords with a score of 2.5 to 5 are considered strong (cf. section 3).

We also investigated the effect of the input modality on password strength and gaze behaviour. We used a repeated-measures ANOVA (with Greenhouse-Geisser correction if sphericity was violated). This was followed by post-hoc pairwise comparisons using Bonferroni-corrected t-tests. Finally, we analysed the post-study questions.

5.2 Feature Extraction

To train the classifiers, we derived a set of seven features that best describe gaze behaviour while entering passwords and are commonly used in literature [Jacob and Karn 2003; Raptis et al. 2017]. For extracting the features, we pre-processed the gaze data. First, we removed irrelevant data. As we used a wearable eye tracker, we also collected gaze data focusing on areas beyond the device screen and keyboard. We only considered gaze data inside the AOI (i.e.., screen and keyboard) and removed the rest. We then identified fixations using the Dispersion-Threshold Identification algorithm [Salvucci and

Goldberg 2000]. It produces accurate results in real-time using only two parameters, that are dispersion and duration threshold (set to 25 and 100, respectively). We then extracted seven low-level gaze features from the defined two areas of interest (AOI), keyboard and screen.

Selected main features used for classification are: 1) avg fixation duration, 2) fixation duration, 3) avg saccadic duration, 4) avg left pupil diameter, 5) avg right pupil diameter, 6) screen fixations count, 7) keyboard fixations count.

In addition to those seven main features we considered the duration spent while typing the password as well as the ratio between the fixations count on the screen and the fixation count on the keyboard. We used thresholding to split the gaze data points between the screen and the keyboard. Differences between AOI are not statistically significant. Hence, we did not take them into further consideration.

5.3 Classification Approach

The goal of our classifier is to map a feature vector computed from a window of data to one of the classes corresponding to the password strength (weak vs strong). We implemented two classifier: a *user-independent*, modality-dependent classifier, trained on the data from different users but using the same input modality and a *user-dependent*, modality-dependent classifier, again using the same input modality. As different classification models generate different levels of performance, we compared three classifiers with a leave one out classification approach: support vector machines (SVM), decision trees, and random forest.

5.3.1 User-independent, Modality-dependent Classifier. We created a user-independent, modality-dependent classifier by training the models on all users for both modalities available (laptop and smartphone). To ensure that the classifiers are performing well on all distributions of data, we split the data into 3 sets: testing, validation, and training. The test set consists of a participant who was not included in training. The validation set was used for tuning the hyper-parameters of the employed machine learning model. It included data of one randomly selected participant with a specified seed for a participant who has been already included in the training set. The password used in the validation set is also not included in the training set. Finally, the training set included the data of all remaining participants. We used a "leave one participant out" cross-validation. For this purpose, we trained and evaluated the classifier for each modality 15 times and each time for a specific participant.

5.3.2 User-dependent, Modality-dependent Classifier. The goal of building user-dependent and modality-dependent classifiers was to determine if better accuracy could be achieved using a personalised model. The classifier was created once for each participant for each input modality. Again, we separated the data into the three sets mentioned above and we used "leave one observation out" cross-validation. For this, we trained and evaluated the classifier 15 times each, using all features, for each participant.

6 RESULTS

6.1 Weak vs Strong Passwords

We collected 366 passwords from all participants in all conditions. In this section, we analyse the passwords collected and report the effect of the different passwords strength on the following:

6.1.1 Passwords Entropy. Table 2 shows the characteristics of the weak and strong passwords used for the comparison, as suggested by [Egelman et al. 2013]. We found that passwords perceived as strong by participants were indeed characterized by a high entropy, i.e. they were indeed considered as actually strong by the password strength estimator.

Table 2. Passwords' characteristics for weak and strong (laptop and smartphone). We compare the password length in characters, number of upper and lower case characters, number of digits, symbols or special characters in the password, whether the password starts with an upper case letter, ends with a lower case letter and finally, we show the zxcvbn strength estimator entropy score

		Password Length	Number of upper case characters	Number of lower case characters	Number of digits in the password	Passwords count that start with upper case characters	Password count that ends with digit	Number of symbols in the password	zxcvbn Entropy score
Weak	**Mean**	7.25	0.41	5.02	1.74	0.21	0.39	0.07	14.61
	SD	3.87	0.92	4.05	2.47	0.41	0.49	0.33	3.59
Strong	**Mean**	15.32	2.25	7.4	3.45	0.49	0.37	1.96	60.76
	SD	6.67	2.22	5.47	2.69	0.5	0.49	2.67	9.20
Laptop	**Mean**	11.17	1.13	6.18	2.93	0.27	0.44	0.82	36.88
	SD	6.63	1.77	4.59	2.65	0.45	0.5	1.87	7.01
Phone	**Mean**	11.01	1.44	6.12	2.19	0.4	0.33	1.11	35.75
	SD	6.76	2	5.17	2.74	0.49	0.47	2.26	8.45

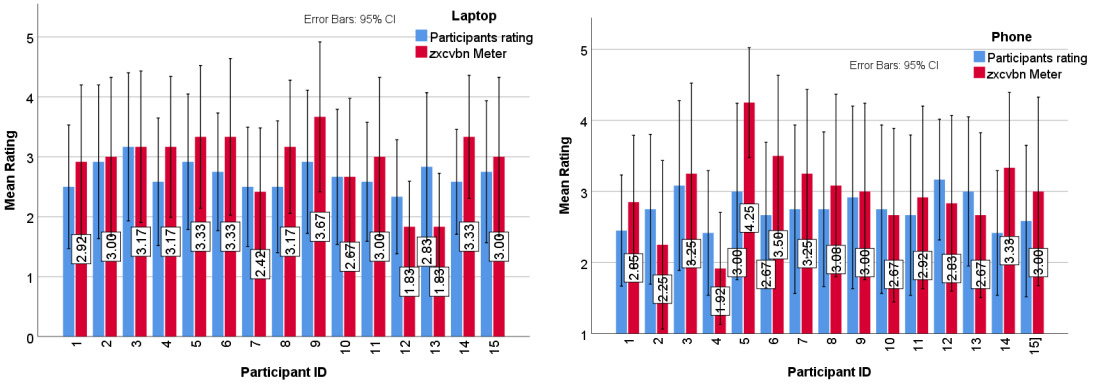

Fig. 2. Laptop (Left) and smartphone (Right) strength comparison between participants' rating and the zxcvbn rating. Showing similar ratings between the zxcvbn meter and users ratings.

This was also reflected in the statistical tests. An ANOVA reveals a statistically significant difference between the entropy of the weak (M = 14.45; SD = 3.59) and the strong passwords (M = 60.75; SD = 9.21), ($F_{1,14}$ = 268.760, $P < .001$). An ANOVA did not show a statistically significant effect for the input device laptop (M = 35.75; SD = 8.45) and smartphone (M = 36.89; SD = 7.02) on the password entropy generated by the zxcvbn password strength estimator, $P > 0.05$. This suggests that the input modality did not affect the generated passwords' actual strength.

6.1.2 Rated Password Strength. To understand how participants perceive their passwords' strength, we compared the users' rated password strength to the strength as indicated by the zxcvbn strength estimator. Figure 2 and 3 compares the average rating for all the passwords entered per participant against the results from the zxcvbn password meter. While there is a variance between passwords ratings, the difference between both ratings is not statistically significant – neither for laptop ($\chi^2(1)$ = 3.769, = .0521) nor for smartphone ($\chi^2(1)$ = 1.66, P = .197), as found by a Friedman test.

The Friedman test also did not reveal a significant effect of the modality on the strength of the weak ($\chi^2(1)$ = 3.6, P = .058) and strong ($\chi^2(1)$ = 3.267, P = .071) passwords. This suggests there might be no difference between perceived and actual password strength. It also shows that the input modality did not affect the strength of the entered password.

In summary, we found that the input modality did not affect the strength or the entropy of the password. This means that participants entered similar passwords on both input modalities. Additionally, we found a statistically significant

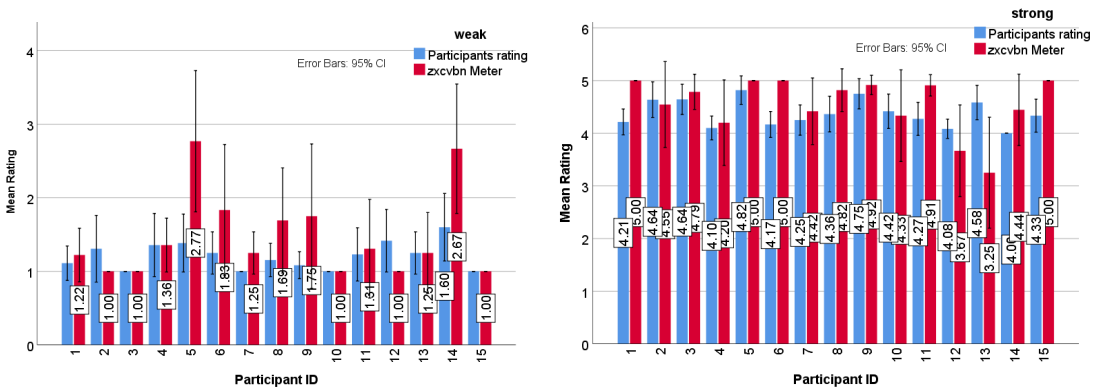

Fig. 3. Weak (Left) and strong (Right) password strength comparison between participants' rating and the zxcvbn password meter rating. Showing similar ratings between the zxcvbn meter and users ratings.

difference between weak and strong passwords' entropy and strength which means participants were able to create password that were rated as weak and strong by the password strength estimator.

6.2 Post Study Questions Analysis

At the end of the study, we asked participants what characteristics makes a strong password. Participants named special characters (22%), adding numbers (18%) upper/lower case characters (18%), and, finally, increasing the length (14%), adding numbers (14%) and adding random characters (14%).

6.3 Gaze Behaviour Statistical Analysis

To assess the relationship between passwords strength and gaze behaviour, we conducted repeated-measures ANOVA.

6.3.1 Effect of Modality on Gaze Behaviour. We tested the effect of the input modality (laptop vs smartphone) on the gaze features (see Table 3). We found that for strong passwords, the input modality has a statistically significant effect on the average fixation duration, fixation duration, average saccadic duration, and keyboard as well as screen fixation count. This means that entering strong passwords on laptops induces shorter fixations, longer saccades, and more fixations on the keyboard as well as less fixations on the screen, compared to smartphone. Participants enter longer passwords on laptops than smartphones. In contrast, for weak passwords, the input modality did not have a strong impact on most gaze data, except for the left pupil diameter[8], screen and keyboard fixation count. This means that entering weak passwords on laptops induces less fixation on the screen and more fixations on the keyboard and also smaller pupil dilation than on smartphones.

6.3.2 Effect of Password Strength on Gaze Behaviour for Input Modalities. To understand the influence of password strength on the gaze features, we ran a repeated-measures ANOVA on the gaze features for both laptops and smartphones. We found that for laptops, entering passwords of different strength has a significant effect on the average fixation duration, fixation duration, average saccadic duration, average left pupil diameter, screen and keyboard fixation count. In particular, entering weak passwords on laptops induces longer fixation duration, shorter saccadic length, smaller left

[8]Possibly due to the *dominant eye effect*. We were not able to verify this as we did not assess participants' dominant eye. We leave this for future work.

Table 3. ANOVA results for eye movements comparing between weak and strong passwords across modalities.(significant in bold)

Eye gaze Feature	Strong Passwords	Pairwise Comp. (Bon.Corr.) (Mean; SD)		Weak Passwords	Pairwise Comp. (Bon.Corr.) (Mean; SD)	
	ANOVA ($F(1,14)$; P)	Laptop	Smartphone	ANOVA ($F(1,14)$; P)	Laptop	Smartphone
Avg fixation dur.	$F = 31.012$; $\mathbf{P < .001}$	.94 ± .017	.96 ± .015	$F = .330$; $P > .05$	.95 ± .015	.95 ± .023
Fixation dur.	$F = 31.012$; $\mathbf{P < .001}$	112.70 ± 2.09	114.61 ± 1.75	$F = .290$; $P > .05$	113.56 ± 1.80	113.81 ± 2.78
Avg saccadic dur.	$F = 31.012$; $\mathbf{P < .001}$	.061 ± .017	.045 ± .015	$F = .290$; $P > .05$	.053 ± .015	.052 ± .023
Avg L pupil diameter	$F = 4.039$; $P > .05$	3.49 ± .38	3.72 ± .54	$F = 12.071$; $\mathbf{P = .004}$	3.39 ± .38	3.69 ± .55
Avg R pupil diameter	$F = .095$; $P > .05$	3.35 ± .59	3.38 ± .89	$F = .625$; $P > .05$	3.25 ± .64	3.35 ± .89
Screen fixation count	$F = 6.377$; $\mathbf{P = .024}$	65.87 ± 22.31	87.27 ± 19.54	$F = 9.246$; $\mathbf{P = .009}$	55.37 ± 22.16	84.14 ± 27.67
Keyboard fixation count	$F = 6.377$; $\mathbf{P = .024}$	54.12 ± 22.31	32.72 ± 19.54	$F = 9.246$; $\mathbf{P = .009}$	64.63 ± 27.16	35.85 ± 27.67
Password duration	$F = 2.14$; $P = 165$	13.5 ± 8.8	11.02 ± 4.3	$F = .056$; $P = .817$	6.5 ± 2.9	6.7 ± 2.8

Table 4. ANOVA results for eye movements comparing between laptop and smartphone during creating weak and strong passwords.

Eye gaze Feature	Laptop	Pairwise Comp. (Bon.Corr.) (Mean; SD)		Smartphone	Pairwise Comp. (Bon.Corr.) (Mean; SD)	
	ANOVA ($F(1,14)$; P)	Strong	Weak	ANOVA ($F(1,14)$; P)	Strong	Weak
Avg fixation dur.	$F = 8.339$; $\mathbf{P = .012}$	.94 ± .017	.94 ± .015	$F = 3.182$; $P > .05$	.96 ± .015	.95 ± .23
Fixation dur.	$F = 8.401$; $\mathbf{P = .012}$	112.68 ± 2.09	113.57 ± 1.80	$F = 3.182$; $P > .05$	114.61 ± 1.75	113.82 ± 2.78
Avg saccadic dur.	$F = 8.401$; $\mathbf{P = .012}$	.060 ± .017	.053 ± .015	$F = 3.182$; $P > .05$	.045 ± .015	.051 ± .023
Avg L pupil diameter	$F = 4.984$; $\mathbf{P = .042}$	3.50 ± .38	3.39 ± .37	$F = .756$; $P > .05$	3.72 ± .54	3.69 ± .55
Avg R pupil diameter	$F = 1.497$; $P > .05$	3.33 ± .59	3.25 ± .69	$F = 1.970$; $P > .05$	3.38 ± .89	3.35 ± .89
Screen fixation count	$F = 6.453$; $\mathbf{P = .024}$	65.87 ± 22.31	55.37 ± 22.16	$F = .847$; $P > .05$	87.28 ± 19.54	84.14 ± 27.68
Keyboard fixation count	$F = 6.453$; $\mathbf{P = .024}$	54.13 ± 22.32	64.63 ± 22.16	$F = .847$; $P > .05$	32.72 ± 19.54	35.86 ± 27.68
Password duration	$F = 12.77$; $\mathbf{P = .003}$	13.5 ± 8.8	6.5 ± 2.9	$F = 25.28$; $\mathbf{P < .001}$	11.02 ± 4.3	6.7 ± 2.8

pupil diameter, fewer fixations on the screen and more fixations of the keyboard. We repeated the same analysis for the smartphone. We did not find a statistically significant effect of the password strength on the gaze behaviour (see Table 4). A reason for this might be that for smartphones gaze is more strongly affected by the area around the device, which might have had an influence on gaze behavior.

6.4 Classifiers Performance

To measure the performance of the classifiers, we computed the Area Under the Curve (AUC), as proposed by Abdelrahman et al.[Abdelrahman et al. 2019]. It aggregates precision and recall into one metric. We also investigated the effect of using user-dependent and user-independent classifiers on the classification of passwords' strength for both modalities.

We first compared the performance of the classifiers on 3 different models: decision trees, random forests, and SVMs. Each classifier was tuned with its relative hyper-parameters to achieve the best results.

As shown in Table 5, the three classifiers resulted in similar AUC, with SVM performed best in most cases. Hence, for the remainder of our analysis, we focus on the SVM results. We found that it is possible to differentiate between strong and weak passwords from users' gaze, independent from the user. The accuracy is 78% for laptops and 76% for smartphones. The user-dependent classifiers outperformed the the user-independent for each modality. They achieve an accuracy of 86% on smartphone and 80% on laptops. We report the true positive and true negative rates using the normalised confusion matrix over all participants for each of the user-independent classifier for both modalities in Figure 4.

6.4.1 Feature Importance. We used the SHAP [Lundberg and Lee 2017] algorithm to investigate the importance of features on the performance of the model for classifying weak and strong password. The SHAP algorithm explains the output of any machine learning model by computing the contribution of each feature to its prediction. The feature

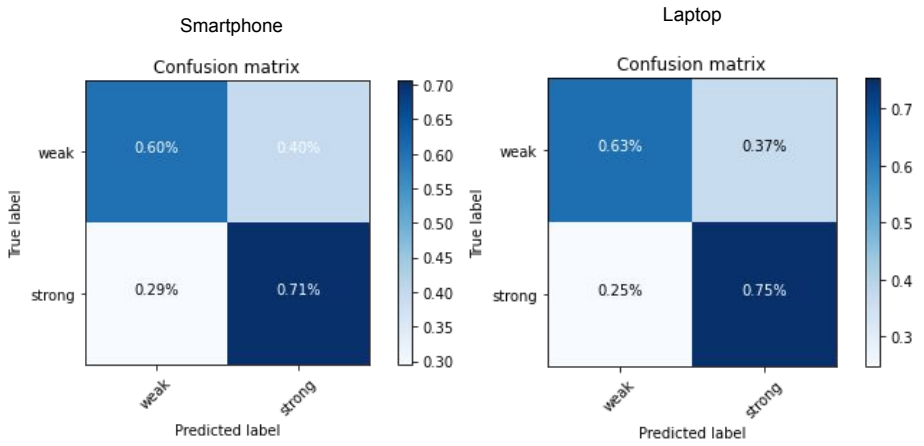

Fig. 4. Confusion matrix for the user-independent, modality dependent classifier for mobile (Left) and laptop (Right).

Table 5. The AUC of the three classification (Decision trees, Random Forests, and SVMs) for smartphone and laptop. The three classifiers have similar accuracy but SVM performs better in most of the results. The best results is highlighted in bold.

	SVM		**Random Forest**		**Decision Tree**	
	Phone	**Laptop**	**Phone**	**Laptop**	**Phone**	**Laptop**
User-indep., Modality-dep.	**.76 ± .19**	.78 ± .16	.76 ± .2	**.79 ± .15**	.70 ± .17	.71 ± .14
User-dep., Modality-dep.	**.86 ± .23**	**.80 ± .29**	.80 ± .25	.71 ± .29	.76 ± .28	.64 ± .29

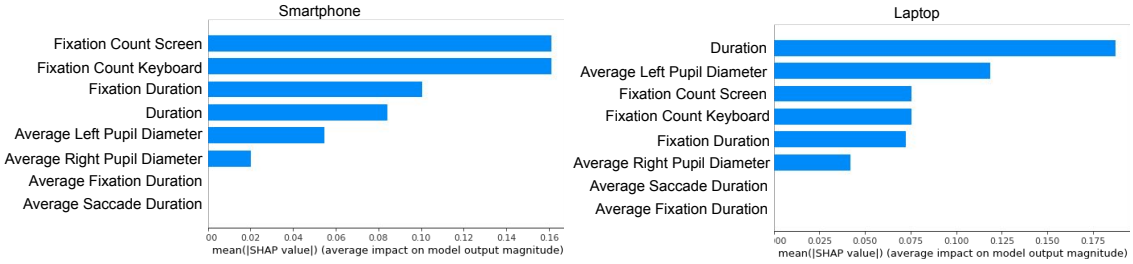

Fig. 5. Features importance for the user-independent, modality dependent classifier for smartphone (Left) and laptop (Right).

importance graph for the SVM model is shown in Figure 5. We observed that for the smartphone modality the fixation count and fixation duration on the smartphone screen and the keyboard are significant in deciding the strength of the password entered by the user. Followed by this, the duration spent while typing the password plays a significant role in the model prediction. For laptops, we observed that the duration has the highest contribution on differentiating between weak and strong passwords followed by the pupil diameter and the fixation count on the screen and keyboard.

6.4.2 Scan Path. Figure 6 shows the different gaze plots for one Participant while creating weak and strong password on both modalities. For laptops strong passwords, participants had more fixations on the screen and keyboard compared to during creating weak passwords. For smartphones, participants had more fixations on the keyboard (area 2) in case of strong passwords compared to weak passwords where they had more fixations on the screen (area 1).

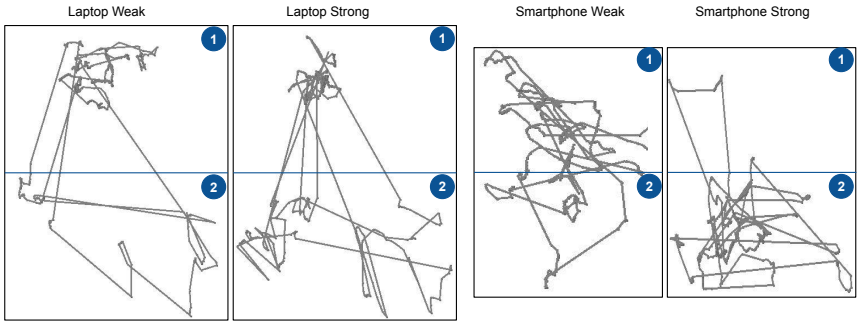

Fig. 6. Gaze Plots, highlighting behaviour while creating weak and strong passwords divided by the areas of interest (1) Screen and (2) Keyboard for both laptops (Left) and smartphones (Right).

7 DISCUSSION

Prior work showed that it is possible to assess graphical password strength based on eye gaze. We applied this idea to detect the strength of text-based passwords from users' gaze behaviour and provided an in-depth investigation. Here, we summarise and discuss the results grouped by different observations.

7.1 Classification Performance

Our results show that password strength classification is feasible, achieving an accuracy of up to 86% when using user-dependent, modality-dependant classifiers. This result is promising as it paves the way for integrating gaze behaviour in authentication where perceived password strength plays an important role, e.g., password strength meters.

When comparing the performance with a user-dependent classifier, we observed a decrease in the accuracy to 76% for smartphones and 74% for laptops. This performance might be sufficient for some applications and is substantially better than guessing. Yet, if high accuracy is crucial future systems might want to employ user-dependent classifiers.

Our results show that it is possible to distinguish the strength of text-based passwords by using gaze features and duration spent while typing the password for user-dependent classification. User-independent results were still strong, suggesting that by training the classifier on one specific task, the classification generalises well to unseen users. This is also confirmed by the statistical analysis of the effect of the input modality and password strength on the gaze features.

It is important to highlight that password characteristics are likely to have an influence on gaze metrics. For example, passwords that include many upper and lower case characters are likely to influence features such as the fixation ratio between keyboad and screen. We only tested with a limited number of users and passwords, so it is likely that such cases are under-represented in our sample. In future work we will investigate the genralsability of our gaze metrics across different password characteristics. In any case, classifiers need to be re-trained for such cases and it is possible that the contribution of features to the classification accuracy might be different.

7.2 Features Performance

The feature importance graph for the SVM classifier shows that the fixation count substantially contributes to the classification accuracy. This is more pronounced for the laptop condition. One explanation for this is that people generally entered longer passwords on the laptop, resulting in this being a more suitable feature. We ran a Pearson correlation between the gaze features and password perceived strength. This, however, did not reveal any statistically

significant effect of the password perceived length on any of the gaze features. We might simply not have had enough data to reveal such a correlation. Apart from this, for both laptops and smartphones pupil dilation is a strong feature. This can be explained through the increased cognitive load while creating strong passwords. In the literature, it has been proven that higher cognitive load leads to an increase in pupil dilation [Duchowski et al. 2018].

7.3 Input Modality Effect

As noticed from our analysis, it was more difficult to estimate password strength from users' gaze behaviour while entering passwords on smartphones compared to laptops. This can be due to the small screen size, as a result of which gaze movements might be more subtle. Also, the way each user holds the phone is different. Some of the users prefer to have the phone closer to their face than other. Besides, some participants used the phone using one hand and others used it with two hands. All of these can be factors that affected classification accuracy. In contrast, on laptops, the distance between the screen and keyboard is larger and, hence, gaze movements are easier to observe. Additionally, we found that participants generate significantly stronger passwords on laptops than on smartphones. This can be due to the different behaviours and reasons behind the use of input modalities. For example, participants might be more used to PINs and lock patterns on smartphones [Harbach et al. 2014; Von Zezschwitz et al. 2013]. In contrast, on laptops users are more likely to authenticate using text-based passwords [Florencio and Herley 2007].

7.4 Influence on Security

Finally, the question arises to which degree the presented approach has an influence on security in general. While our approach is primarily meant to be used by researchers and practitioners to design novel approaches that ultimately lead to stronger passwords, knowledge on password strength in the hand of an attacker might have an adverse effect. For example, if an attacker gets access to an eye tracker, they might find out which users employ weaker passwords of for which accounts they employ weaker passwords, making those a more likely target of an attack.

8 CONCLUSION

We introduced a novel approach of using gaze behaviour as an additional metric to assess password strength. Our approach assesses users' gaze behaviour while creating passwords. We hypothesised that the way in which users create strong and weak passwords is reflected in their gaze behavior. Our results confirmed our hypothesis and showed that it is possible to differentiate between weak and strong passwords with an accuracy of 86% for personalised classifiers on smartphone and 80% on laptops. Our findings pave the way for using gaze behaviour in security interfaces, in particular interfaces that make people use stronger passwords.

Future work could collect datasets that focus on different password characteristics, settings, input modalities as well as user characteristics (e.g., dominant eye) to investigate for which cases the approach generalises well. Another interesting aspect is the influence of password reuse on the approach. Also, trying to classify password strength in a more fine-grained manner could be interesting. Finally, future work could look into novel concepts. In particular, we see potential in approaches that are independent of the input device.

ACKNOWLEDGMENTS

This work was supported by the Royal Society of Edinburgh (RSE award no. 65040), EPSRC New Investigator Award (EP/V008870/1), DFG grant no. 316457582 and 425869382, dtec.bw-Digitalization and Technology Research Center of the Bundeswehr (Voice of Wisdom), and finally, the Studienstiftung des deutschen Volkes.

REFERENCES

Yomna Abdelrahman, Anam Ahmad Khan, Joshua Newn, Eduardo Velloso, Sherine Ashraf Safwat, James Bailey, Andreas Bulling, Frank Vetere, and Albrecht Schmidt. 2019. Classifying Attention Types with Thermal Imaging and Eye Tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 69 (Sept. 2019), 27 pages. https://doi.org/10.1145/3351227

Yasmeen Abdrabou, Mohamed Khamis, Rana Mohamed Eisa, Sherif Ismail, and Amrl Elmougy. 2019. Just Gaze and Wave: Exploring the Use of Gaze and Gestures for Shoulder-Surfing Resilient Authentication. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications* (Denver, Colorado) *(ETRA '19)*. Association for Computing Machinery, New York, NY, USA, Article 29, 10 pages. https://doi.org/10.1145/3314111.3319837

E. R. Abdulin and O. V. Komogortsev. 2015. Person verification via eye movement-driven text reading model. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. 1–8.

Akram Bayat and Marc Pomplun. 2018. Biometric Identification Through Eye-Movement Patterns. In *Advances in Human Factors in Simulation and Modeling*, Daniel N. Cassenti (Ed.). Springer International Publishing, Cham, 583–594.

Joseph Bonneau, Cormac Herley, Paul C. Van Oorschot, and Frank Stajano. 2015. Passwords and the evolution of imperfect authentication. *Commun. ACM* 58, 7 (1 July 2015), 78–87. https://doi.org/10.1145/2699390 Copyright: Copyright 2018 Elsevier B.V., All rights reserved.

A. Bulling, J. A. Ward, H. Gellersen, and G. TrÃűster. 2011. Eye Movement Analysis for Activity Recognition Using Electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 4 (2011), 741–753.

Virginio Cantoni, Chiara Galdi, Michele Nappi, Marco Porta, and Daniel Riccio. 2015. GANT: Gaze analysis technique for human identification. *Pattern Recognition* 48, 4 (2015), 1027–1038.

Virginio Cantoni, Tomas Lacovara, Marco Porta, and Haochen Wang. 2018. A Study on Gaze-Controlled PIN Input with Biometric Data Analysis. In *Proceedings of the 19th International Conference on Computer Systems and Technologies* (Ruse, Bulgaria) *(CompSysTech'18)*. Association for Computing Machinery, New York, NY, USA, 99–103. https://doi.org/10.1145/3274005.3274029

Anupam Das, Joseph Bonneau, Matthew Caesar, Nikita Borisov, and XiaoFeng Wang. 2014. The tangled web of password reuse.. In *NDSS*, Vol. 14. 23–26.

Xavier de Carné de Carnavalet and Mohammad Mannan. 2014. From very weak to very strong: Analyzing password-strength meters. In *Network and Distributed System Security Symposium (NDSS 2014)*. Internet Society.

Andrew T. Duchowski, Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos. 2018. The Index of Pupillary Activity: Measuring Cognitive Load Vis-à-Vis Task Difficulty with Pupil Oscillation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173856

M. Dupuis and F. Khan. 2018. Effects of peer feedback on password strength. In *2018 APWG Symposium on Electronic Crime Research (eCrime)*. 1–9.

Serge Egelman, Andreas Sotirakopoulos, Ildar Muslukhov, Konstantin Beznosov, and Cormac Herley. 2013. Does My Password Go up to Eleven? The Impact of Password Meters on Password Selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2379–2388. https://doi.org/10.1145/2470654.2481329

Dinei Florencio and Cormac Herley. 2007. A Large-Scale Study of Web Password Habits. In *Proceedings of the 16th International Conference on World Wide Web* (Banff, Alberta, Canada) *(WWW '07)*. Association for Computing Machinery, New York, NY, USA, 657–666. https://doi.org/10.1145/1242572.1242661

Dinei Florêncio, Cormac Herley, and Paul C. van Oorschot. 2014. Password Portfolios and the Finite-Effort User: Sustainably Managing Large Numbers of Accounts. In *23rd USENIX Security Symposium (USENIX Security 14)*. USENIX Association, San Diego, CA, 575–590. https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/florencio

Alain Forget, Sonia Chiasson, and Robert Biddle. 2010. Shoulder-Surfing Resistance with Eye-Gaze Entry in Cued-Recall Graphical Passwords. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1107–1110. https://doi.org/10.1145/1753326.1753491

Alain Forget, Sonia Chiasson, P. C. van Oorschot, and Robert Biddle. 2008. Persuasion for Stronger Passwords: Motivation and Pilot Study. In *Persuasive Technology*, Harri Oinas-Kukkonen, Per Hasle, Marja Harjumaa, Katarina Segerståhl, and Peter Øhrstrøm (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 140–150.

Marian Harbach, Emanuel Von Zezschwitz, Andreas Fichtner, Alexander De Luca, and Matthew Smith. 2014. It's a Hard Lock Life: A Field Study of Smartphone (Un)Locking Behavior and Risk Perception. In *Proceedings of the Tenth USENIX Conference on Usable Privacy and Security* (Menlo Park, CA) *(SOUPS '14)*. USENIX Association, USA, 213–230.

John M Henderson, Svetlana V Shinkareva, Jing Wang, Steven G Luke, and Jenn Olejarczyk. 2013. Predicting cognitive state from eye movements. *PloS one* 8, 5 (2013), e64937.

Sabrina Hoppe, Tobias Loetscher, Stephanie A. Morey, and Andreas Bulling. 2018. Eye Movements During Everyday Behavior Predict Personality Traits. *Frontiers in Human Neuroscience* 12 (2018), 105. https://doi.org/10.3389/fnhum.2018.00105

Robert JK Jacob and Keith S Karn. 2003. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye*. Elsevier, 573–605.

Christina Katsini, Yasmeen Abdrabou, George E. Raptis, Mohamed Khamis, and Florian Alt. 2020. The Role of Eye Gaze in Security and Privacy Applications: Survey and Future HCI Research Directions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3313831.3376840

Christina Katsini, Christos Fidas, George E. Raptis, Marios Belk, George Samaras, and Nikolaos Avouris. 2018a. Influences of Human Cognition and Visual Behavior on Password Strength during Picture Password Composition. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173661

Christina Katsini, George E. Raptis, Christos Fidas, and Nikolaos Avouris. 2018b. Towards Gaze-Based Quantification of the Security of Graphical Authentication Schemes. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications* (Warsaw, Poland) *(ETRA '18)*. Association for Computing Machinery, New York, NY, USA, Article 17, 5 pages. https://doi.org/10.1145/3204493.3204589

Mohamed Khamis, Florian Alt, and Andreas Bulling. 2018. The Past, Present, and Future of Gaze-enabled Handheld Mobile Devices: Survey and Lessons Learned. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Barcelona, Spain) *(MobileHCI '18)*. ACM, New York, NY, USA. https://doi.org/10.1145/3229434.3229452

Mohamed Khamis, Florian Alt, Mariam Hassib, Emanuel von Zezschwitz, Regina Hasholzner, and Andreas Bulling. 2016. GazeTouchPass: Multimodal Authentication Using Gaze and Touch on Mobile Devices. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 2156–2164. https://doi.org/10.1145/2851581.2892314

Mohamed Khamis, Mariam Hassib, Emanuel von Zezschwitz, Andreas Bulling, and Florian Alt. 2017. GazeTouchPIN: Protecting Sensitive Data on Mobile Devices Using Secure Multimodal Authentication. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (Glasgow, UK) *(ICMI '17)*. Association for Computing Machinery, New York, NY, USA, 446–450. https://doi.org/10.1145/3136755.3136809

Manu Kumar, Tal Garfinkel, Dan Boneh, and Terry Winograd. 2007. Reducing Shoulder-Surfing by Using Gaze-Based Password Entry. In *Proceedings of the 3rd Symposium on Usable Privacy and Security* (Pittsburgh, Pennsylvania, USA) *(SOUPS '07)*. Association for Computing Machinery, New York, NY, USA, 13–19. https://doi.org/10.1145/1280680.1280683

Michael D Leonhard and VN Venkatakrishnan. 2007. A comparative study of three random password generators. In *2007 IEEE International Conference on Electro/Information Technology*. IEEE, 227–232.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*. 4765–4774.

Päivi Majaranta and Kari-Jouko Räihä. 2007. Text entry by gaze: Utilizing eye-tracking. *Text entry systems: Mobility, accessibility, universality* (2007), 175–187.

Gilbert Notoatmodjo and Clark Thomborson. 2009. Passwords and perceptions. In *Proceedings of the Seventh Australasian Conference on Information Security-Volume 98*. Citeseer, 71–78.

George E. Raptis, Christina Katsini, Marios Belk, Christos Fidas, George Samaras, and Nikolaos Avouris. 2017. Using Eye Gaze Data and Visual Activities to Infer Human Cognitive Styles: Method and Feasibility Studies. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (Bratislava, Slovakia) *(UMAP '17)*. Association for Computing Machinery, New York, NY, USA, 164–173. https://doi.org/10.1145/3079628.3079690

Caitlin Rinn, Kathryn Summers, Emily Rhodes, Joël Virothaisakun, and Dana Chisnell. 2015. Password creation strategies across high-and low-literacy web users. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–9.

Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying Fixations and Saccades in Eye-Tracking Protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (Palm Beach Gardens, Florida, USA) *(ETRA âĂŹ00)*. Association for Computing Machinery, New York, NY, USA, 71–78. https://doi.org/10.1145/355017.355028

Tobias Seitz, Manuel Hartmann, Jakob Pfab, and Samuel Souque. 2017. Do Differences in Password Policies Prevent Password Reuse?. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 2056–2063. https://doi.org/10.1145/3027063.3053100

Richard Shay, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Alain Forget, Saranga Komanduri, Michelle L. Mazurek, William Melicher, Sean M. Segreti, and Blase Ur. 2015. A Spoonful of Sugar? The Impact of Guidance and Feedback on Password-Creation Behavior. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2903–2912. https://doi.org/10.1145/2702123.2702586

Richard Shay, Saranga Komanduri, Adam L. Durity, Phillip (Seyoung) Huh, Michelle L. Mazurek, Sean M. Segreti, Blase Ur, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2014. Can Long Passwords Be Secure and Usable?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 2927–2936. https://doi.org/10.1145/2556288.2557377

Elizabeth Stobert and Robert Biddle. 2016. Expert Password Management. In *Technology and Practice of Passwords*, Frank Stajano, Stig F. Mjølsnes, Graeme Jenkinson, and Per Thorsheim (Eds.). Springer International Publishing, Cham, 3–20.

Blase Ur, Felicia Alfieri, Maung Aung, Lujo Bauer, Nicolas Christin, Jessica Colnago, Lorrie Faith Cranor, Henry Dixon, Pardis Emami Naeini, Hana Habib, Noah Johnson, and William Melicher. 2017. Design and Evaluation of a Data-Driven Password Meter. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3775–3786. https://doi.org/10.1145/3025453.3026050

Blase Ur, Jonathan Bees, Sean M. Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2016. Do Users' Perceptions of Password Security Match Reality?. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3748–3760. https://doi.org/10.1145/2858036.2858546

Blase Ur, Patrick Gage Kelley, Saranga Komanduri, Joel Lee, Michael Maass, Michelle L. Mazurek, Timothy Passaro, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2012. How Does Your Password Measure Up? The Effect of Strength Meters on Password

Creation. In *21st USENIX Security Symposium (USENIX Security 12)*. USENIX Association, Bellevue, WA, 65–80. https://www.usenix.org/conference/usenixsecurity12/technical-sessions/presentation/ur

Blase Ur, Fumiko Noma, Jonathan Bees, Sean M. Segreti, Richard Shay, Lujo Bauer, Nicolas Christin, and Lorrie Faith 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)*. USENIX Association, Ottawa, 123–140. https://www.usenix.org/conference/soups2015/proceedings/presentation/ur

D. Vitonis and D. W. Hansen. 2014. Person Identification Using Eye Movements and Post Saccadic Oscillations. In *2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems*. 580–583.

Emanuel Von Zezschwitz, Paul Dunphy, and Alexander De Luca. 2013. Patterns in the wild: a field study of the usability of pattern and pin-based authentication on mobile devices. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services*. 261–270.

Ding Wang and Ping Wang. 2015. The emperorâĂŹs new password creation policies. In *European Symposium on Research in Computer Security*. Springer, 456–477.

Matt Weir, Sudhir Aggarwal, Michael Collins, and Henry Stern. 2010. Testing Metrics for Password Creation Policies by Attacking Large Sets of Revealed Passwords. In *Proceedings of the 17th ACM Conference on Computer and Communications Security* (Chicago, Illinois, USA) *(CCS '10)*. Association for Computing Machinery, New York, NY, USA, 162–175. https://doi.org/10.1145/1866307.1866327

Daniel Lowe Wheeler. 2016. zxcvbn: Low-budget password strength estimation. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*. 157–173.

Yongtuo Zhang, Wen Hu, Weitao Xu, Chun Tung Chou, and Jiankun Hu. 2018. Continuous Authentication Using Eye Movement Response of Implicit Visual Stimuli. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 177 (Jan. 2018), 22 pages. https://doi.org/10.1145/3161410