Authentication Schemes - Comparison and Effective Password Spaces

Peter Mayer[1], Melanie Volkamer[1], and Michaela Kauer[2]

[1] Center for Advanced Security Research Darmstadt, Technische Universität Darmstadt, Germany
[2] Institute of Ergonomics, Technische Universität Darmstadt, Germany

Abstract. Text passwords are ubiquitous in authentication. Despite this ubiquity, they have been the target of much criticism. One alternative to the pure recall text passwords are graphical authentication schemes. The different proposed schemes harness the vast visual memory of the human brain and exploit cued-recall as well as recognition in addition to pure recall. While graphical authentication in general is promising, basic research is required to better understand which schemes are most appropriate for which scenario (incl. security model and frequency of usage). This paper presents a comparative study in which all schemes are configured to the same effective password space (as used by large Internet companies). The experiment includes both, cued-recall-based and recognition-based schemes. The results demonstrate that recognition-based schemes have the upper hand in terms of effectiveness and cued-recall-based schemes in terms of efficiency. Thus, depending on the scenario one or the other approach is more appropriate. Both types of schemes have lower reset rates than text passwords which might be of interest in scenarios with limited support capacities.

Keywords: Usable Security, Authentication, Graphical Passwords

1 Introduction

Text passwords are the most common means of authentication. Despite this ubiquity, they have been the target of much criticism. User-created passwords are highly predictable. Most users compose their passwords solely of lower case characters, use simple dictionary words or put numbers and special characters at easy foreseeable places [16, 28]. Furthermore, users have on average 25 accounts, but only seven passwords [16, 19]. This password reuse raises serious concerns when considering that many websites transmit and store passwords in the clear instead of encrypted and cryptographically hashed [3]. Password managers are in many cases also no solution: They introduce a single point of failure; the security of password managers depends on the strength of the master password; and there are portability issues.

These deficits of text passwords motivated many researches to find alternatives. One alternative is graphical authentication. Like text passwords, graphical

authentication schemes are knowledge-based. Their primary goal is to exploit the vast visual memory of the human brain. Visual memory is superior to memory of abstract information such as text [23]. Many different graphical authentication schemes have been proposed and studies have been conducted to assess their security and usability (e.g. [6, 8, 35]). While graphical authentication is in general promising there are also drawbacks like efficiency when authenticating. Therefore, it is important to understand how different schemes perform wrt. to usability (including efficiency, effectiveness, satisfaction but also memorability) and security in comparison to each other. Most past studies only provide information in comparison to text passwords; and as the experimental settings differ from study to study, this data cannot be used to compare the different schemes and approaches to each other. The authors of prior comparative studies either studied only cued-recall-based or recognition base schemes. Also, to our knowledge, most studies base their configurations of the graphical authentication schemes on the theoretical password space. This is a severe limitation of such studies and renders a comparison an impossible task, because alternative schemes that force or persuade users to choose more secure (and therefore potentially less memorable) passwords are compared side by side with schemes that let the users choose their passwords freely (and therefore potentially very insecurely).

Therefore, more basic research is necessary to enable the comparison and to support decision makers in selecting the most appropriate authentication scheme for their scenario. In this paper we present the first usability study of multiple graphical authentication schemes and text passwords that uses the most recent literature available on the effective password space of the tested schemes as baseline for the security configuration. Furthermore, this study is among the first to compare schemes based on recognition and schemes based on cued-recall in the same experimental setting. The selected graphical schemes are: PassPoints, PCCP, Faces and Things. Participants were asked to login five times over a period of 42 days. 337 participants took part. The evaluated usability measures are derived from the measures used in prior literature and therefore allow a comparison to existing research. The results of the experiment are:

- Usability-wise: The results demonstrate that recognition-based schemes have the upper hand in terms of effectiveness and cued-recall-based schemes in terms of efficiency. We also show that with only one exception the graphical schemes in our study have significantly lower reset rates than text passwords. In addition, we found evidence that male participants like the graphical schemes better after longer times of usage, while female participants find text passwords easier to use. Yet, female participants are more willing to use graphical password schemes than male participants.
- Security-wise: The analysis of the actual password space shows how difficult the prediction of effective password spaces is and that further research is necessary to better judge on the security level of some schemes. The estimates from our study can serve as baseline for configurations in future studies.

– Comparison to prior studies: Our study provides evidence that no significant quantitative difference between the performance of male and female participants exist, as the results of prior studies could not be replicated. However, significant differences in the attitude towards the schemes exist.

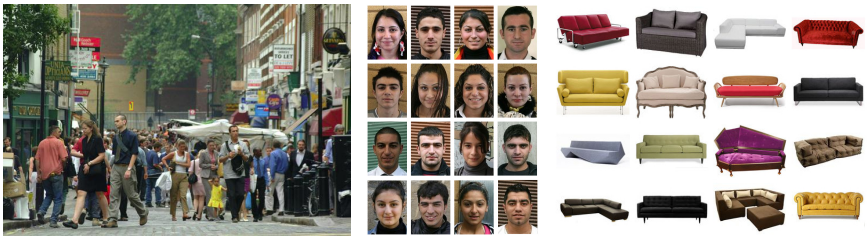
2 The password space

The password space of an authentication scheme is the set of all passwords and therefore closely related to the guessing resistance of the scheme. The larger the password space, the more guesses (on average) are necessary to find the right password. However, it is important to distinguish between the theoretical and the effective password space. While the theoretical password space includes all possibly selectable passwords for a scheme, the effective password space comprises only the subset of passwords which are likely to be actually chosen. Often the effective and theoretical password spaces are different. The only definitive exception in this regard are system assigned random passwords. Yet, these are not applicable in many scenarios as such passwords are usually more difficult to remember. Assessing the effective password space can be difficult, because a sufficiently large sample of passwords is needed to derive any meaningful information on frequently appearing values [21].

Multiple metrics have been proposed to compare and predict the password spaces of different schemes. The measure most often used to assess the size of the effective password space is the Shannon entropy of recorded password samples [28]. It is also the basis for the recommendations in the NIST Electronic Authentication Guideline and has been used in research of text passwords (e.g. [22, 28]) as well as graphical passwords (e.g. [9]). However, it has been found, that neither the NIST estimates nor Shannon entropy provide truely reliable estimates and represent more a "rule of thumb" than an accurate metric, especially since sample sizes in typical usability studies are far smaller than what would be desirable [21, 33]. Kelley et al. [21] proposed *guess-number calculator* and Bonneau [2] proposed α-*guesswork* as more robust and reliable metrics, but for none of these two empirical values for graphical passwords are available. Thus, in the the absence of viable alternatives, Shannon entropy is used as measure to configure the schemes in this study.

3 Graphical passwords

Like text passwords, graphical passwords are knowledge-based authentication. Today's research distinguishes between three types of graphical password schemes, named after the way they strain the users' memory: purely recall-based schemes, cued-recall-based schemes, and recognition-based schemes. We decided against including purely recall-based schemes like *Draw-A-Secret* (DAS) [20] in our study, since such schemes have been shown to be insecure [24]. In the following we briefly describe the cued-recall-based and recognition-based schemes we included in the study.



(a) PassPoints and PCCP (b) Faces (c) Things

Fig. 1: The interfaces of PassPoints and PCCP (marker of the click point not visible during normal operation), Faces, and Things.

3.1 Cued-recall-based schemes

Graphical authentication schemes based on cued-recall use the graphical elements as cues to support the users' recall. Wiedenbeck et al. proposed *PassPoints* [35] whose basic working principle is the definition of click-points on an image. The image only serves as a cue for the user, the actual password is composed of the coordinates of the click-points. During authentication, the same image as during enrollment is displayed and the user has to select her/his click-points in the right order. Even with the image as cue, perfect cursor positioning cannot be expected from the user. Therefore, a tolerance margin around each click-point compensates small imprecisions by the user.

To avoid click-point patterns in PassPoints, Chiasson et al. proposed an improvement called *Cued Click-Points* (CCP) [4]. Instead of multiple click-points on one image, in CCP users create multiple click-points, each on a different image. During authentication one image is shown after the other and depending on where on the image the user clicks, either the correct next image of her/his password is shown (when the right click-point was selected) or an image for which s/he has not created a click-point before (when the input was incorrect). In the latter case, authentication has to be finished regardless, but the legitimate user can detect her/his error. To counter so called hot-spots (points that have been noticed to be chosen significantly more often than other click-points), Chiasson et al. proposed *Persuasive Cued Click-Points* (PCCP) [4]. In the authentication phase, they employ persuasive technology by means of an additional viewport during enrollment. While the system offers the possibility to shuffle the viewport, the authors found that the number of hot-spots is significantly reduced and the distribution of click-points in PCCP does not significantly differ from a random distribution [4]. PassPoints and PCCP have an extensive coverage in research literature enabling estimates for the effective password space. Therefore, the two schemes were selected for the usability study. Figure 1(a) shows the interface of the two schemes.

3.2 Recognition-based schemes

In recognition-based authentication schemes users need only to recognize their password among a variety of choices, which substantially decreases the required mental effort. *Passfaces* is the most popular recognition-based scheme [26] and is commercially available. During enrollment, the system assigns the users several facial images as their passwords. During authentication, as many grids of facial images as there are faces in the user's password are displayed one after another. In each grid, the user has to identify the image belonging to her/his password. Thus the whole authentication consists of multiple rounds, one for each face in the password.

The same technique as used in Faces was also attempted with objects other than faces [11, 18]. Due to its strong coverage in the literature we included a Passfaces-like scheme in the study, which is subsequently referred to as *Faces* (see Fig. 1(b)); as well as a scheme using objects, which is subsequently referred to as *Things* (see Fig. 1(c)).

4 Related work

In the following we briefly discuss selected related work. Chiasson et al. present in [4] a comparison of the three graphical cued-recall-based authentication schemes PCCP, CCP and PassPoints. They cover usability and security aspects with results from eight user studies. Their results indicate no significant differences in success rates between the three schemes. The timings they recorded are 6-8 seconds for PassPoints and 8-15 seconds for PCCP. The configurations in all studies are based on the theoretical password space.

Hlywa et al. present a comparison of multiple Passfaces-like recognition-based graphical authentication schemes in two web-studies [18] in which recognition-based schemes are configured to password-level security. The schemes used in the study of Hlywa et al. differ only in the type of images. One scheme used faces (similarly to Faces), one scheme used everyday objects (similarly to Things) and one scheme used houses. Both of their studies used configurations similar, but not identical, to those in our experiment. One study used 4 by 4 grids, but instead of a password length of 7 rounds Hlywa et al. used 5. Their other study used a different approach to the grid size, but was configured to use the same password space, namely 2^{28}. In both studies system-assigned passwords were used. They report significantly faster login times for their objects participants than for their faces participants in both studies.

Stobert and Biddle compared in [29] the performance of text passwords as well as recall-based and recognition-based graphical passwords. Their results indicate, that recognition-based passwords had a higher memorability, but their usability was limited by longer login times. Our methodology differs from the one employed by Stobert and Biddle in two key aspects. Firstly, we consider already proposed schemes (some of them used in the wild), while Stobert and Biddle specifically designed a new scheme to compare the different types of memory

Scheme	Configuration	$\mathbf{H}_{exp}$
Text	8 characters length, policy based on basic8survey [22]	~ 27.19
PassPoints	4 click-points, 600x400 px resolution of image, 9x9 px tolerance margin [9]	$\lesssim 26$
PCCP	3 click-points, 570x380 px resolution of images, 19x19 px tolerance margin	~ 27.7
Faces	7 grids of 16 images each, facial images	$\sim$28
Things	7 grids of 16 images each, object images, semantically grouped	$\sim$28

Table 1: Configurations (H_{exp} = expected Shannon entropy in bits)

retrieval. Secondly, we base the security configurations of the schemes we use on the effective password space, rather than the theoretical password space.

Schaub et al. analysed and described the design space for graphical passwords on smartphones [27]. To test their design metric, they conducted a non-longitudinal lab-based user study with five different graphical password schemes and PINs. Their schemes' configurations are based on the theoretical password space rather than the effective password space. Also, their usability metrics use different definitions. Therefore, the comparability to this study is very limited.

5 Configuration of the schemes

For the configuration of the schemes we decided to focus solely on the effective password space. Other attacks such as spyware, shoulder surfing and social engineering, (e.g. mentioned by [31]) were not considered in our study.

Most major websites use password policies resulting in a minimum of about 20 to 28 bits of entropy (or password spaces of 2^{20} to 2^{28} respectively) [17]. Thus, configuring the authentication schemes' effective password spaces to lie within these two values is a close approximation to a lower bound to what would actually be seen in the wild. However, we wanted to narrow down the range for our study. To do so, we analysed Komanduri et al. research on password policies for text passwords and their resulting effective password spaces [22]. They present only one policy whose effective password space lies between 20 and 28 bits of entropy, namely their *basic8survey* policy, for which they report an entropy of 27.19 bits. Therefore, we decided to use this value as baseline and configure the password spaces of all schemes to lie between 2^{27} and 2^{28} or as closely to that target as possible. Table 1 lists all schemes with their configurations. A justification for each scheme is given below in the respective subsections.

5.1 Text password

The basic8survey policy of Komanduri et al. only states "Password must have at least 8 characters" and does not offer any scenario to the participants [22]. For this policy they reported 27.19 bits of entropy. We used this policy as starting

point. We slightly modified it due to some design considerations: Most importantly we made the policy look more like as if it could be from one of the major websites investigated in [17] and additionally enforced a fixed length of 8 characters. Our password policy was included in the introductory text for the scheme and stated the following rules:

1. The passwords are case sensitive. For instance, "EXAMPLE" and "Example" are two different passwords. When choosing a password, remember the capitalization you use.
2. The password must be exactly 8 characters long.
3. Try to not use passwords that are easy to guess, for example "password".
4. Even though it is not a requirement, try to add a number or special character to your password.
5. Choose a new password, that you are not using for any other service.

The first statement is merely a further explanation of the scheme and not a policy rule, it is often found in policies on the Internet. The first actual policy rule regards the password length and is taken from Komanduri et al. [22], but was changed to a fixed length of eight characters. This was done as all graphical schemes can only be configured in such a restricted way. Rules 3 and 4 were added as incentive for the participants to not create "just study passwords" whose main purpose was to be not forgotten. This decision was made in conjunction with the decision to include priming in the study methodology. The last rule was intended to motivate the participants to deal with a new password throughout the study, just as the participants of the other groups would have to. Additionally, the system enforced to choose passwords not mentioned in the policy (e.g. "password" as mentioned in the third rule could not be chosen).

5.2 Cued-recall-based schemes

The alphabet A of the cued-recall-based schemes is the number of distinct click-points on the image. This number is determined by the size of the image divided by the tolerance margin. The theoretical password space P can be calculated from the number of click-points c required by the system: $P = |A|^c$.

PassPoints. The empirical entropy estimates of Dirik et al. in [9] seem to be the only findings for PassPoints passwords allowing an estimation of the effective password space. Therefore, we decided to adopt their PassPoints configuration and use an image with a resolution of 400x600 px. However, we follow prior studies, which argue to use 9 px for the tolerance margin instead of 10 px as an uneven number can be centered around one pixel while it is is unclear how this should be done for an even margin [4]. Regarding the image, we decided to use the people image used by Dirik et al. as the sole image for our PassPoints implementation, as it scored best in their evaluation. They report an observed entropy of 6.5 bits for a click-point on that image. The authors remark that these values are only valid for multiple click-points if independence between the click-points in the password is assumed and that the actual entropy should

be assumed to be lower. Seemingly no data on the magnitude of dependence between click-points in PassPoints passwords is available in published literature. Therefore, we decided to use four click-points, as the resulting entropy value of 26 bits (password space of 2^{26} respectively) is closer to the target than any other number of click-points.

PCCP. According to Chiasson et al. the distribution of click-points in PCCP does not significantly differ from a random distribution [5]. Therefore, it can be assumed that the theoretical and the effective password space are approximately identical. We decided to use three click-points on 570x380 px images with a 19 px tolerance margin for the PCCP implementation, which results in an expected entropy of 27.7 bits and thus meets the entropy target. The tolerance margin of 19 px was chosen, as this value achieved better usability ratings than viable alternatives [34]. The image size was chosen as a multiple of the tolerance margin and determined in conjunction with the number of click-points to model the target entropy values of 27 to 28 bits as good as possible. Priority was given to image size, as it was shown that larger images lead to less clustering in PCCP [30]. As the images used in previous studies by Chiasson et al. could not be made available for our study due to copyright concerns, all images for the implementation were taken from the same source as the image for PassPoints. The viewport size of 75 px was adopted from [5]. We decided to restrict user choice by omitting the possibility to shuffle the viewport for the following reason: Chiasson et al. report that most participants either do not shuffle at all or they shuffle a lot in order to circumvent the persuasion mechanism. The latter case negates the security advantage of PCCP over PassPoints.

5.3 Recognition-based schemes

Recognition-based schemes suffer from predictability issues if users are allowed to choose their passwords themselves [7]. Consequently, commercial implementations assign random passwords to the users. Furthermore, to our knowledge, no entropy estimates for the configurations of Things and Faces are available. Thus, we decided to assign random passwords to the participants for both recognition-based schemes as recommended by [10, 13, 18]. Then, the theoretical and effective password space P from the alphabet A (comprised of the images in one grid) and the number of required authentication rounds r is: $P = |A|^r$. To reach the entropy target of 27 to 28 bits, we decided to use a 4x4 grid and 7 rounds, i.e. a password space of $16^7 = 2^{28}$.

Faces. All images used were taken from the Face of Tomorrow project which were also used by other studies regarding facial recognition in authentication [13, 18]. We ensured that all grids had images with similar backgrounds and that faces used in one grid were of people from the same ethnicity.

Things. The Things scheme used semantically grouped images, following the work of Weinshall and Kirkpatrick [32]. They found that pictures should not be too similar to one another or users would start confusing them and consequently advise to select "pictures with a clear central subject or action and [...] differences within the group" [32]. Therefore, this criteria was applied to

all images gathered for the Things scheme. All images were informally reviewed to fulfill this criteria by colleagues uninvolved in the selection.

6 User study methodology

Due to the high relevancy of web-authentication, it was decided to conduct an online web-based study. Each participant had to complete five sessions over a period of about 42 days. The intervals between the sessions prolonged, namely they were: 1 day, 3 days, 7 days, 30 days. The first part of each session comprised the interaction of the participants with their assigned authentication scheme: either creation or authentication. Creation was divided into five phases: (1) the participants entered their user name, (2) an introductory text was shown, (3) the participants created their passwords or the system-assigned passwords were generated, (4) participants could review their new passwords, and (5) a short training to familiarize the participants with their new scheme and to confirm the password. Authentication had two phases: first the participants entered their user name, then they entered their password. A short questionnaire, investigating the participants' impressions regarding the usability and security of their assigned scheme, concluded each session. The participants had a 24 hour window to complete each session. Participants were informed of their sessions by email and reminders were sent out if they had only 5 hours left in their 24 hour window. If at any point during authentication or creation a participant could not remember her/his password, s/he could reset it via an automated procedure on the website. To motivate participation throughout the whole study, a raffle was held for all participants who completed all five sessions. The study was available in English and German. The methodology of this usability study conforms to all requirements of our university's ethics committee. Only the data relevant to our analysis was recorded and participants could request deletion of their data at any time. No additional scenario was presented to the participants, they were fully primed in terms of the password schemes. Fahl et al. have shown that this procedure is scientifically sound [14]. Additionally, Bonneau found that security motivations such as registered payment information has no greater impact than demographic factors [2].

6.1 Participants

Participants were assigned to their schemes using stratified sampling of three factors: the participants sex, the language the participant enrolled in and a 5-point Likert value of the participant's self-assessed experience in password security. The participant's sex was chosen as factor due to previous literature suggesting differences in performance between male and female users [6]. The language the participant enrolled in was chosen as a factor to prevent bias originating from the translation of the questionnaires as well as the instructions explaining the study's procedure and the operation of the authentication schemes to the participants. The self-assessed experience in password security was chosen as a factor

due to an evaluation of questions regarding the participants' real life passwords, which is not part of this paper.

Participants were recruited internationally in various ways, including, but not limited to, flyers and posters on campus, mailing lists, forums and social networks. All participants had to enroll on the study's website using a registration form. No paid services such as Amazon's Mechanical Turk or CrowdFlower were used. Overall 337 participants registered for our study and confirmed their email address. Of those 250 were male and 87 female. Participants registered in both languages available in the study, namely 262 participants used German texts and 75 used English texts. The age range was 14 to 67 years (mean = 27.6, median = 25). Almost half the participants (46.5%) reported to have experience in the field of password security (scores of 4 or 5 on a 5-point Likert scale).

6.2 Recorded measures

The recorded usability measures are aligned along the ISO 9241-11 criteria effectiveness, efficiency and satisfaction.

Effectiveness. The first measure regarding effectiveness is the *success rate*. To retain the highest degree of comparability to past research, we follow the best practices of Biddle et al. in [1] and report success rates after the first and after the third attempt. Additionally, this study includes overall success rates (no limit on the attempts), as this is the most frequently reported measure in the literature. In this study success rates are defined as follows: the rate of participants having successfully authenticated after a certain amount of attempts (one, three, no limit) to the total number of participants using the scheme. The second effectiveness measure is the password *reset rate*. It describes the average number of resets per participant which is a more comparable measure than absolute values due to different numbers of participants assigned to the different schemes. We also report *dropout rates*, which is the number of participants dropped out in relation to the total number of participants assigned to a scheme. Note that the dropout is a rather unreliable measure of effectiveness, especially considering the narrow 24 hours time frame participants had to complete each session.

Efficiency. The first efficiency measure recorded during our study is the *interaction time* in seconds. It only counts the time for the actual interaction with the system and not the time the participants' browsers need to load the images. This prevents unpredictable bias due to different Internet connection speeds. The second measure of efficiency is the *number of attempts needed*, which represents the efficiency measure for each session. This measure complements the interaction times, which are the efficiency measure for each attempt. In order to assess the overall efficiency of the schemes, the two aforementioned measures (interaction times and number of attempts) are combined to calculate *expected average total authentication times*. This measure is what comes closest to traditional authentication timings which measure the time for the overall login procedure (from login request to completed login). However, it has to be stressed that this is an approximation of what is to be expected for the average participant. Yet, it offers a much better comparability between the schemes than the traditionally

reported timings, which often depend on the schemes' implementations (i.e. loading times, etc.). The *time needed to read the instructions* is the total time across all sessions to account for participants who had to reset their password and read through the instructions again. These times are also reported in seconds. This measure serves as an indicator regarding the learnability of the schemes. In order to spot implementation issues and gather information on possibly necessary improvements of the schemes' implementations, the analysis also examines *system times*. These predominantly include the time the system needs to load the necessary contents (in particular the images) from the study server. The system times complement the interaction times to detect usability issues caused by the technical side of our implementations.

Satisfaction. *Questionnaires* at the end of each session captured the participants' satisfaction with their assigned scheme. The questions concerned the participants' attitude and impressions regarding the usability and security of their scheme. The majority is implemented using a 5-point Likert scale (5 represents strong agreement and 1 represents strong disagreement).

7 Results

In the following we describe the results of our usability study. The first section concerns itself with the validity of the security assumption (i.e. whether the target entropy values were reached). Then we present the usability results along the lines of the criteria effectiveness, efficiency and satisfaction of ISO 9241-11.

7.1 Validity of the security assumptions

In order to check the validity of our assumptions regarding the effective password spaces, we calculated empirical entropy values for all schemes. As we already explained in section 2, due to the small password samples the estimates reported below should only be seen as approximations of the actual differences in the effective password spaces of the schemes. The values of the recognition-based schemes, whose passwords were randomly assigned to the participants, can serve as an indication regarding this deviation. Table 2 shows for each scheme the expected entropy values derived from published research and the empirical values calculated for the passwords of our study.

Text passwords. Table 2 shows the entropy estimates calculated from the study text passwords according to [28] and the value reported by Komanduri et al. for their *basic8survey* policy. With an entropy of 27.41 bits the text passwords created by the participants of our study are very close to the target of 27.19 bits.

PassPoints. Table 2 lists the entropy values calculated from the PassPoints passwords according to Dirik et al. in [9] as well. Entropy values are calculated for each click-point position in the password (analogously to the character positioning in text passwords). These values come very close to those of Dirik et al. but undercut them. Also, the value is calculated under the assumption that the choice of all click-points is independent from the other click-points chosen in

	Target	Calculated
Text	27.19	27.41
PassPoints (upper bound)	26.00	24.17
PCCP	27.69	17.35
Faces	28.00	26.52
Things	28.00	26.79

Table 2: Target and empirical entropy values in bits

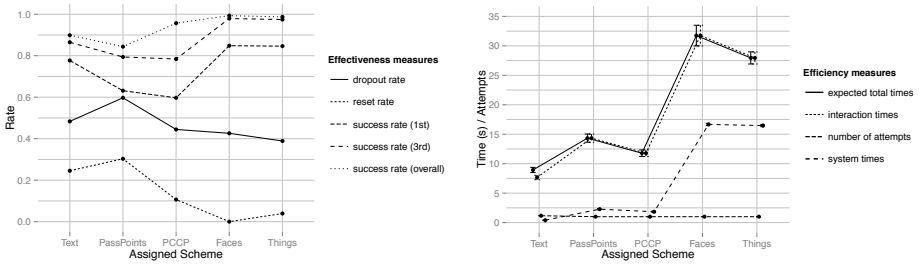
the same password. This is not a reasonable assumption [9], but still the target entropy is not reached.

PCCP. The calculated empirical entropy of the PCCP passwords in our study is 17.35 bits and thus far below the target value of 27.69 bits. This result deviates considerably from what was expected.

7.2 Usability Evaluation

The following three sections present the results of the usability evaluation. Most of our data is not normally distributed and/or has heterogenous variances. Therefore, we use robust alternatives to standard tests as suggested by Field [15], Erceg-Hurn and Mirosevich [12] and Wilcox [36] in favor of data transformations. In detail, the used tests were Fisher's Exact Test (FET), the ANOVA-type statistic tests (ATS) developed by Brunner et al. and implemented in the R packages `WRS` by Wilcox and `nparLD` by Noguchi et al. [25] as well as Cliff's test (Cliff) as described by Wilcox in [36]. All multiple comparisons use Holm-Bonferroni corrected α-levels. Also, in order to cope with the outliers in the timing data (where participants would leave the session open and return after some time to continue the session), we used robust measures of location in favor of the mean. The standard errors are also calculated with regards to these robust measures of location. Readers unfamiliar with these statistical methods can find further information in appendix A if desired. All factors and their interactions not mentioned in the results were analyzed, but left out since they did not show significant results. Figure 2 shows all effectiveness measures and all efficiency measures. These aspects are discussed in more detail in the following paragraphs.

Effectiveness. The effect of the assigned scheme on the success rates is significant after the first attempt (FET: $p < .001$). Significant differences occur for the pairs Text-PassPoints (FET: $p = .006$), Text-PCCP (FET: $p. = 001$), PassPoints-Faces (FET: $p < .001$), PassPoints-Things (FET: $p < .001$), PCCP-Faces (FET: $p < .001$) and PCCP-Things (FET: $p < .001$). From the results of follow-up Fisher's tests, a bipartition after the first attempt becomes apparent: PassPoints and PCCP (lower group) display both significantly worse success rates than Text, Faces and Things (upper group). After three attempts the effect of the assigned scheme on the success rates is again significant (FET: $p < .001$). The bipartition mostly remains, only the Text group moves somewhat



(a) Effectiveness measures

(b) Efficiency measures

Fig. 2: Measured Results

between the upper and the lower group. In detail, significant differences are found for the following pairs: Text-Faces (FET: $p < .001$), Text-Things (FET: $p < .001$), PassPoints-Faces (FET: $p < .001$), PassPoints-Things (FET: $p < .001$), PCCP-Faces (FET: $p < .001$) and PCCP-Things (FET: $p < .001$). For the overall success rates (no limit on the attempts considered), the effect of the assigned scheme on the success rates is also significant (FET: $p < .001$). However, the bipartition is lost. PassPoints scores lowest with significant differences to all schemes except Text. The Text group also shows significant differences to the recognition-based schemes: Text-Faces (FET: $p < .001$) and Text-Things (FET: $p < .001$). In the overall scores, the recognition-based schemes show the best scores. PCCP scores only non-significantly worse. The Text scheme shows significant differences to the recognition-based-schemes, but not to PCCP. Detailed success rates for each session after one attempt, three attempts and overall can be found in appendix B for closer inspection.

As becomes apparent from the reset rates depicted in Fig. 2(a), the variation in the number of password resets is very large: the reset rates vary from 0.3 to 0. A Fisher's test shows that the effect of the assigned scheme is highly significant (FET: $p < .001$). Upon examining the scores of the schemes, a partitioning in three groups becomes apparent. Text and PassPoints show the highest reset rates (0.25 and 0.30). PCCP displays a rate of 0.11 and the recognition-based schemes show the best rates with 0.00 (Faces) and 0.04 (Things).

Figure 2(a) also shows dropout rates for the schemes. The scheme with the highest dropout rate is PassPoints (60%), the one with the lowest is Things (39%). The differences in dropout between the schemes are however not significant. The most important reason for dropout is the 24 hours time frame participants had to return to the study website and complete their session. 53.7% of all dropout can be attributed to this time frame.

Efficiency. From the interaction times plotted in Fig. 2(b) it becomes clear, that for recognition-based schemes it takes much longer to enter the password. A three-way ATS test shows significant main effects for the assigned scheme ($H(4.32, 43.24) = 55.59$, $p < .001$) and the session ($H(2.82) = 5.05$,$p < .001$). For the assigned scheme, the only non-significant difference is between Faces and

Things. The Text group shows the lowest interaction times. The second lowest score is the one of the PCCP group, then follows PassPoints and last are the two recognition based schemes, whose interaction times were twice to three times as high. Regarding the sessions, the only significant differences are between session 2 and 4 and between session 4 and 5. Participants needed longer in the second session, than in the third and fourth, but need the most time in the fifth session. A table with a more detailed breakdown of the interaction times of each session can be found in appendix B.

The second efficiency indicator beside the interaction times is the number of attempts needed. The main effects of the assigned scheme ($H(3.90, 29.27) = 4.98$, $p = .003$) and the sessions ($H(2.35) = 5.34$, $p = .002$) are significant in an ATS test. The values for the different schemes are depicted in Fig. 2(b). For the assigned scheme, the significant differences occur for the pairs Text-PCCP (Cliff: $p = .004$) and Text-Faces (Cliff: $p = .002$). Concerning the sessions, participants needed more login attempts in the later sessions (4 and 5) than in the earlier sessions. The significant differences for the sessions occur between sessions 2 and 5 ($H(1) = 12.25$, $p < .001$), between sessions 3 and 5 ($H(1) = 14.19$, $p < .001$) and between sessions 4 and 5 ($H(1) = 14.27$, $p < .001$).

While no total authentication times were recorded, expected average total authentication times can be approximated from the average number of authentication attempts and the average interaction times. The two factors "assigned scheme" and "session" are considered in this calculation. As in neither of the two relevant analyses the participants' sex shows a significant effect, its influence can be neglected. The resulting times are depicted in Fig. 2(b) alongside the interaction times. A table with all the values in detail can be found in appendix B. The advantages and disadvantages of some schemes in both analyses more or less annihilate. It is still expected for users of the recognition-based schemes to take twice as long as users of other schemes. However, e.g. the advantage of the Text group over PCCP in the interaction times is eaten up by the higher number of attempts needed in the later sessions. However, this measure should only be seen as an approximation and therefore be treated with caution.

Female participants took on average significantly more time to read the instructions (123.27 sec) than males (97.22 sec) did ($H = 968.71$, $p = .029$). This is the only measure for which the participants' sex had a significant main effect. The effect of the assigned scheme is non-significant, as is the interaction. This indicates no significant difference in the learnability between all schemes.

Even if interaction times are low, participants might discard a scheme as unusable, if the system itself takes too long to respond. Figure 2(b) shows the system times for all schemes. An ATS test shows a significant result for the assigned scheme. In fact, all differences except the one between Faces and Things are significant.

Satisfaction. The participants' attitude towards the system was investigated using questionnaires. Figure 3 shows a summary of the average answers to the following 5-point Likert questions: (Q1) The password scheme is easy to use, (Q2) I could remember my password easily, (Q3) Entering my password was

fast, (Q4) The creation of my password was easy, (Q5) I think I can remember my password easily, (Q6) I think I can remember my password more easily than passwords I normally use, (Q7) I prefer this new password scheme to my previous passwords, and (Q8) I would recommend this password scheme to others.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	
Text	4.25	3.78	4.02	3.71	3.53	2.61	2.45	3.10	1
PassPoints	3.92	4.08	3.28	3.71	3.50	2.90	2.52	3.03	2
PCCP	4.27	4.03	3.74	3.70	3.14	3.33	2.97	3.87	3
Faces	4.29	4.06	3.44	4.00	3.65	3.42	2.77	3.32	4
Things	4.55	4.17	3.77	4.02	3.67	3.39	3.15	3.85	5
Male	4.30	4.05	3.74	3.93	3.56	3.24	2.74	3.38	
Female	4.26	4.08	3.80	3.74	3.58	3.18	3.18	3.87	

Fig. 3: The average answers to a variety of 5-point Likert questions.

Participants generally found their assigned authentication scheme easy to use (Q1) and easy to remember (Q2,Q5), but not easier to remember than their current passwords (Q6). PCCP was perceived significantly more difficult to remember ($H(4.68, 97.25) = 3.21$, $p = .012$). Perception of how fast password entry is differs depending on the assigned scheme (Q3). PassPoints scores lowest. The Text group perceives its scheme as being the fastest. Faces, Things and PCCP are rated equally. Despite the low efficiency apparent from the interaction times, the recognition-based schemes are still perceived to be faster than PassPoints. These differences are significant ($H(3.47, 20.45) = 3.43$, $p = .031$). Overall, only Things scores favorably in terms of a change towards it, though none of the differences are significant (Q7). While the perceived gain in usability does not seem to be large enough for the participants to happily adopt the new systems, they would generally recommend them to others (Q8).

The analysis of question Q2 also revealed a significant interaction of all three factors (assigned scheme, participants' sex and session). While text passwords always score lowest with the male participants, its rating from the female participants steadily increases up to the point where it scores highest in the last session. Faces preserves its ratings from the male participants over the course of the sessions and scores the highest rating in the fifth session, but is rated lowest by the female participants in all sessions except the fourth where it scores second to last. Thus, the scheme female participants perceive to have the highest memorability in the final session is rated lowest by the male participants and the scheme rated highest by male participants is rated lowest by female participants. These differences seem however to to be purely subjective as they are not mirrored by the actual performance of the participants.

8 Discussion

Configuring the different authentication schemes to the same effective password space is an important aspect of this study's methodology. This goal could only

be partially attained. Table 2 shows the target entropy values and the entropy values calculated for the study passwords. For the Text scheme the target has been reached with only a negligible discrepancy. The entropy for the cued-recall-based schemes is lower than expected. The PassPoints entropy estimates for each position in the password (6.15, 6.04, 5.74 and 6.28 bits) are lower than the 6.5 bits found by Dirik et al. in [9]. The individual estimates for PCCP (5.70, 5.85 and 5.80 bits) are even lower than those for PassPoints. The difference in image resolution and tolerance margin size are crucial in this regard. However, such entropy estimates have, to our knowledge, never been reported for PCCP in published literature. Thus, a precise prediction of this loss in entropy was hard if not impossible, especially when considering that user choice was further restricted by omitting the shuffle mechanism. Chiasson et al. stated that there are "no significant differences between PCCP and what is expected to occur by chance" [30]. More research has to be conducted in order to find the relation between their finding and the difference between effective password space and theoretical password space discovered in this study.

The security analysis shows how difficult reliable estimation of the effective password space is. However, empirical determination is the only way to gather reliable data in this regard. The Shannon entropy values reported in this study can provide estimates for future studies. Yet, whenever possible improved metrics such as Bonneau's α-guesswork should be used in favor of Shannon entropy.

The implications of the differences regarding the entropy among the schemes are unclear. At best the differences are small enough to have no effect, although this seems unlikely. At worst, the differences are the same as in other studies, relying on the theoretical password space without any regard to the actual entropy. It is important to keep these findings in mind when interpreting the usability results. Though it has to be mentioned that due to the very small samples we had, all empirically calculated metrics should only be seen as an approximation. This becomes especially clear when regarding the entropy estimates of the recognition-based schemes, which should converge towards 28 bits of entropy in larger samples, due to their random nature.

In the usability evaluation, no scheme emerges as the sole victor of our comparison. The recognition-based schemes have higher effectiveness ratings than all other schemes. This advantage is interesting, as participants of Faces and Things were assigned random passwords and the entropy of the passwords was higher than for the cued-recall-based passwords. The cued-recall-based schemes have better efficiency ratings, where the conceptually more complex scheme PCCP scores even better than its predecessor PassPoints, especially in the later sessions. The interaction times are the only measure in which the Text group scores best. Text passwords scored worse than PCCP in terms of expected average timings when intervals between logins were long, but this might be attributed to the higher entropy in the text passwords.

The satisfaction ratings are mostly similar for all schemes, with two exceptions. Firstly, PassPoints is perceived to be slower than the recognition-based schemes, despite its better efficiency scores. This emphasizes the usability prob-

lems participants had with the scheme. Secondly, when participants were asked whether they preferred their new scheme, all schemes except Things score on average below 3 on a 5-point Likert scale. This notably includes the Text scheme. It seems that participants are not very fond of their current text passwords, but the alternative schemes offered to them in this study also do not represent their first choice. Thus, further alternatives need to be evaluated in order to identify the most suitable candidate.

One scheme clearly performs worst in basically all aspects: PassPoints. The adjustment of the security-level according to the findings of Dirik et al. results in low usability ratings overall. Consequently, it is safe to say that PassPoints is unsuitable to function as password scheme on a relevant security-level and can be excluded from future studies. For a meaningful comparison between the remaining schemes and especially when considering one of the schemes for actual implementation in a production environment, the requirements of such an environment play the most important role in judging the scheme's suitability.

The password reset rate of the graphical schemes is significantly lower. Therefore, in situations where recovery of lost passwords is expensive in time or effort, graphical passwords seem to be the better choice. If the user tries to authenticate using her/his text password at least once, then decides to reset the password and the reset procedure takes only about 20 seconds (twice the time of a normal login), even recognition-based graphical schemes have the potential to be more efficient than text passwords due to their superior memorability. However, more research is needed to provide conclusive evidence in this regard, since no reset timings were recorded in this study. For the recognition-based schemes only two resets occurred in the course of this study and success rates are generally higher, coming close to 100% even after longer periods of inactivity. Therefore, they are best suited for applications in which logins are infrequent with long intervals between them. In situations with more frequent logins, PCCP might be the better choice. Success rates are equal to those of traditional text passwords, interaction times are only somewhat elevated and the advantage in password resets remains. The aspect of increased memorability becomes ever more relevant when considering the login policies motivated by the online tracking efforts of companies such as Google or Facebook. The expiration dates in the session cookies of these popular services are usually set years into the future, so users will remain logged in as long as possible. Therefore, the logins of users potentially occur very infrequently on the same device.

Throughout the whole analysis we found no evidence for quantitative differences in the performance of male and female participants. Thus, the results of earlier studies by Chiasson et al. [6] could not be replicated. The only aspect where differences between male and female performance could be found is password creation. Male participants need fewer rounds in the training while spending less time to read the scheme instructions and find creation easier. However, in opposition to what is stated by Chiasson et al., this difference in performance seems not to be dependent on the type of mental work (visual-spatial tasks vs linguistic tasks) as for none of these measures a significant interaction between

the assigned scheme and the participants' sex can be found. Yet, the subjective differences in user satisfaction are present and might hinder adoption of graphical passwords especially by male users.

Despite the overall success of this study, it has some limitations that need to be addressed. To configure the effective password space we only considered Shannon entropy. This is one of the most severe limitations of this study. Yet, we argue that it is a necessary compromise we had to make in our first step away from user studies regarding only the theoretical password space. While not optimal, it was the only viable option to approximate the effective password space. For future studies more reliable metrics such as α-guesswork have been proposed in recent literature and should be used instead of Shannon entropy, whenever empirical values allowing estimation of the effective password space are available. Reporting these metrics was unfeasible due to the (for password research) still small sample sizes in this study, but is planned for future work.

A restriction we imposed due to basic design decisions is the fixed length of the passwords for all schemes. For the graphical password schemes this decision was necessary in order to control the theoretical password space. For the text passwords we added the same limitation in order to negate any discrepancies in the usage of the different schemes. However, participants perceive this restriction as unnatural and many wanted more flexibility. However, such flexibility can only be incorporated in a methodology aiming at the same effective password space if studies such as those conducted by Komanduri et al. [22] are available for all schemes in a comparative study. This is obviously not the case for this study, but is an important field for future work.

Also, the decision of comparing schemes with system-assigned passwords to schemes which allow user choice might be considered a limitation of this study. For all the schemes we tried to use implementations, as they might be used for real-world applications. We followed the canonical implementations used in recent studies and distributed as commercial products (user-choice in cued-recall-based schemes and system-assigned passwords in recognition-based schemes).

Last but not least, it has to be noted that the methodology of this study neglects one usability aspect, namely interference of multiple passwords. While it has been shown that its influence is significant, it has been excluded in this study, mainly due to time and recruitment constraints. Future studies should optimally include this aspect.

9 Conclusion

This study compares four graphical authentication schemes and text passwords in a user web-study. It is the first study of its kind to base the security configuration of the tested schemes on the effective password space. An analysis of the security assumptions once again shows how difficult the prediction of effective password spaces is. The estimates reported in this work can serve as baseline for configurations in future studies and represent an individual research contribution.

The results of the usability evaluation show that no scheme emerges as the sole victor of this evaluation, but that all have strengths and weaknesses. Yet, both types of graphical authentication schemes show fewer password resets than the text passwords and in situations were these are costly graphical authentication seems to be the better choice. This is the first time such findings are reported not only inside one category of graphical passwords, but across the cued-recall-based and the recognition-based categories.

The results of Chiasson et al. indicating differences between male and female performance in the usage of graphical password schemes could not be replicated in this study. However, this study found significant differences in the attitude of male and female participants. Additional research is needed in order to find more evidence regarding this issue.

References

1. Biddle, R., Chiasson, S., van Oorschot, P.C.: Graphical passwords: Learning from the first twelve years. CSUR 44(4) (Aug 2012)
2. Bonneau, J.: The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. Proc. IEEE S&P pp. 538–552 (2012)
3. Bonneau, J., Preibusch, S.: The password thicket: technical and market failures in human authentication on the web. Proc. WEIS '10 (Jun 2010)
4. Chiasson, S., Stobert, E., Forget, A., Biddle, R., van Oorschot, P.C.: Persuasive Cued Click-Points: Design, Implementation, and Evaluation of a Knowledge-Based Authentication Mechanism. IEEE Trans. on Dep. and Sec. Comp. 9(2), 222–235 (2012)
5. Chiasson, S., Forget, A., Biddle, R., van Oorschot, P.C.: Influencing users towards better passwords: persuasive cued click-points. In: Proc. BCS-HCI '08 (Sep 2008)
6. Chiasson, S., Forget, A., Stobert, E., van Oorschot, P.C., Biddle, R.: Multiple password interference in text passwords and click-based graphical passwords. In: Proc. CCS '09. pp. 500–511. ACM (Nov 2009)
7. Davis, D., Monrose, F., Reiter, M.K.: On user choice in graphical password schemes. In: Proc. USENIX '04. pp. 151–164 (2004)
8. Dhamija, R., Perrig, A.: Deja Vu: A user study using images for authentication. In: Proc. SSYM '00. pp. 45–58 (2000)
9. Dirik, A.E., Memon, N., Birget, J.C.: Modeling user choice in the PassPoints graphical password scheme. In: Proc. SOUPS '07. pp. 20–28 (2007)
10. Dunphy, P., Yan, J.: Is FacePIN secure and usable? In: Proc. SOUPS '07 (Jul 2007)
11. Ellis, H.D.: Recognizing Faces. Brit. J. of Psychology 66(4), 409–426 (Apr 2011)
12. Erceg-Hurn, D.M., Mirosevich, V.M.: Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. American Psychologist 63(7), 591–601 (2008)
13. Everitt, K.M., Bragin, T., Fogarty, J., Kohno, T.: A comprehensive study of frequency, interference, and training of multiple graphical passwords. In: Proc. CHI '09. pp. 889–898 (2009)
14. Fahl, S., Harbach, M., Acar, Y., Smith, M.: On the ecological validity of a password study. In: Proc. SOUPS '13. pp. 13:1–13:13 (2013)
15. Field, A., Miles, J., Field, Z.: Discovering Statistics Using R. SAGE Publications Limited (Mar 2012)

16. Florêncio, D., Herley, C.: A large-scale study of web password habits. In: Proc. WWW '07. pp. 657–666 (2007)
17. Florêncio, D., Herley, C.: Where do security policies come from? In: Proc. SOUPS '10 (2010)
18. Hlywa, M., Biddle, R., Patrick, A.S.: Facing the facts about image type in recognition-based graphical passwords. In: Proc. ACSAC '11. pp. 149–158 (2011)
19. Ives, B., Walsh, K.R., Schneider, H.: The domino effect of password reuse. Comm. of the ACM 47(4), 75–78 (2004)
20. Jermyn, I., Mayer, A., Monrose, F., Reiter, M.K., Rubin, A.D.: The design and analysis of graphical passwords. Proc. SSYM'99 (1999)
21. Kelley, P.G., Komanduri, S., Mazurek, M.L., Shay, R., Vidas, T., Bauer, L., Christin, N., Cranor, L.F., Lopez, J.: Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. Proc. IEEE S&P pp. 523–537 (2012)
22. Komanduri, S., Shay, R., Kelley, P.G., Mazurek, M.L., Bauer, L., Christin, N., Cranor, L.F., Egelman, S.: Of Passwords and People: Measuring the Effect of Password-Composition Policies. In: Proc. CHI '11. pp. 2595–2604 (2011)
23. Mulhall, E.F.: Experimental Studies in Recall and Recognition. Am. J. of Psych. 26(2), 217–228 (Apr 1915)
24. Nali, D., Thorpe, J.: Analyzing user choice in graphical passwords. School of Comp. Sci. (2004)
25. Noguchi, K., Gel, Y.R., Brunner, E.: nparLD: An R Software Package for the Nonparametric Analysis of Longitudinal Data in Factorial Experiments. J. of Statistical Software 50(12) (Sep 2012)
26. Real User Corporation: The Science Behind Passfaces (Jul 2004)
27. Schaub, F., Walch, M., Könings, B., Weber, M.: Exploring The Design Space of Graphical Passwords on Smartphones. In: Proc. SOUPS '13. ACM (Jul 2013)
28. Shay, R., Komanduri, S., Kelley, P.G., Leon, P.G., Mazurek, M.L., Bauer, L., Christin, N., Cranor, L.F.: Encountering Stronger Password Requirements: User Attitudes and Behaviors. In: Proc. SOUPS '10 (Jul 2010)
29. Stobert, E., Biddle, R.: Memory retrieval and graphical passwords. In: Proc. SOUPS '13. ACM Press, New York, New York, USA (2013)
30. Stobert, E., Forget, A., Chiasson, S., van Oorschot, P.C., Biddle, R.: Exploring Usability Effects of Increasing Security in Click-based Graphical Passwords. In: Proc. ACSAC '10. pp. 79–88 (2010)
31. Suo, X., Zhu, Y., Owen, G.S.: Graphical Passwords: A Survey. In: Proc. ACSAC '05 (2005)
32. Weinshall, D., Kirkpatrick, S.: Passwords you'll never forget, but can't recall. In: CHI EA '04. pp. 1399–1402 (2004)
33. Weir, M., Aggarwal, S., Collins, M., Stern, H.: Testing Metrics for Password Creation Policies by Attacking Large Sets of Revealed Passwords. In: Proc. CCS '10. pp. 162–175 (2010)
34. Wiedenbeck, S., Waters, J., Birget, J.C., Brodskiy, A., Memon, N.: Authentication Using Graphical Passwords: Effects of Tolerance and Image Choice. In: Proc. SOUPS '05. pp. 1–12. ACM (2005)
35. Wiedenbeck, S., Waters, J., Birget, J.C., Brodskiy, A., Memon, N.: PassPoints: Design and longitudinal evaluation of a graphical password system. Int. J. of Hum.-Comp. Studies 63(1-2), 102–127 (Jul 2005)
36. Wilcox, R.R.: Introduction to Robust Estimation & Hypothesis Testing. Elsevier Academic Press, 3rd edition edn. (Feb 2012)

A Statistical methods

The web-based nature of the study imposes some limitations on the data analysis. Some timing values were highly elevated due to participants probably starting the session and then being distracted by other tasks. Therefore, all timings are analysed using medians, modified one-step M estimators (MOM) or trimmed means instead of means. The modified one-step M estimator of location is a measure of central tendency. It accounts for outliers which are determined using the median absolute deviation. A detailed description of the modified one-step M estimator is beyond the scope of this work, but Wilcox and Keselman wrote an excellent introduction to measures of central tendency[1].

All data collected throughout the study turned out to be non-normally distributed. Consequently, traditional parametric statistical procedures could not be applied as even small deviations can cause a severe drop in power of F-statistic tests [36]. The widely used non-parametric Mann-Whitney and Kruskal-Wallis tests were not sufficient for the three-factor design of this study (assigned scheme, sex of the participant and session). Furthermore, even for lower factor analyses they can prove unsatisfactory. The Mann-Whitney test can only control the Type I error rate under the assumption, that the groups have identical distributions [36]. The Kruskal-Wallis test performs only well, if its null hypothesis is true. Otherwise its statistical power is uncertain[2].

Following the advice of Erceg-Hurn and Mirosevich [12] and Wilcox [36], the following statistical methods are applied to scale data. In two-way designs, where the data has no longitudinal factor (e.g. time taken to read instructions), Wilcox advises a two-way ANOVA using modified one-step M estimators and a percentile bootstrap. He provides the R procedure `m2way` to conduct the analysis. Using a similar method is also suggested by Field [15]. Any follow-up tests are conducted using a similar method with a percentile bootstrap and the MOM for two groups (pbMOM). For all three-factor analyses the rank-based ATS described below is used.

All ordinal data is analysed using rank-based methods. This type of data comprises all Likert-scale questions. As outlined above classical rank-based methods are unsatisfactory. Therefore, instead of the Mann-Whitney test Cliff's test is used for the analysis. It performs well with small samples, can handle tied values and has a slight advantage over alternatives if many tied values appear [36]. Wilcox provides the R procedure `cidv2` to conduct this test. Following the recommendations of Erceg-Hurn and Mirosevich in [12], the Kruskal-Wallis test is substituted by the rank-based ANOVA-type statistic (ATS), which allows for heteroscedastic data and tied values [36]. Wilcox provides the R procedure `bdm` to conduct such a test. The non-longitudinal data with two independent variables is analysed using the R procedure `bdm2way` provided by Wilcox, analogously to the

[1] Wilcox, R.R., Keselman, H.J.: Modern Robust Data Analysis Methods: Measures of Central Tendency. Psychological Methods 8(3), 254274 (2003)

[2] Wilcox, R.R.: Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction. CRC Press (Jul 2011)

procedure `bdm`. Follow-up analyses are conducted using Cliff's test. Wilcox provides no ATS procedures for longitudinal data. However, Noguchi et al. adapted the ATS method and published their `nparLD` package for R in 2012 [25]. All longitudinal data is analysed using their procedures.

All tests for independence were conducted using Fisher's exact test due to the robustness in cases with small contingency values. All multiple comparisons use Bonferroni-Holm corrected α-levels. The p-values are given to the third decimal place, if a p-value would be rounded to a value not representable with three decimal places the term $p < .001$ is used.

B Detailed effectiveness and efficiency values

Table 3 gives a more detailed breakdown of the success rates for each scheme over the course of the five sessions. The influence of the sessions become apparent. The most obvious example are the values for PCCP in the fifth session, where the success rate after the first attempt is about half of the overall success rate.

	S 2			**S 3**			**S 4**			**S 5**		
	1st	3rd	∞	1st	3rd	∞	1st	3rd	∞	1st	3rd	∞
Text	0.74	0.86	0.86	0.85	0.92	0.95	0.86	0.94	0.97	0.65	0.71	0.81
PassPoints	0.55	0.82	0.86	0.73	0.83	0.88	0.71	0.87	0.95	0.53	0.63	0.66
PCCP	0.59	0.76	0.93	0.65	0.85	1.00	0.65	0.82	0.93	0.47	0.70	0.90
Faces	0.86	0.98	0.98	0.84	0.97	1.00	0.86	1.00	1.00	0.83	0.97	1.00
Things	0.93	0.98	1.00	0.83	1.00	1.00	0.87	1.00	1.00	0.73	0.91	0.94
Male	0.76	0.90	0.94	0.80	0.94	0.97	0.80	0.93	0.98	0.65	0.79	0.86
Female	0.70	0.88	0.92	0.79	0.88	0.96	0.78	0.92	0.98	0.61	0.80	0.91

Table 3: The authentication success rates for each session after the first attempt (1st), after the third attempt (3rd) and overall (∞); sessions S2-S5.

Table 4 lists all the interaction times and expected average total authentication times. The advantage of the Text group over PCCP in the interaction times is eaten up by the higher number of attempts needed in the later sessions.

	Interaction times				**Expected average**				
	S 2	S 3	S 4	S 5	S 2	S 3	S 4	S 5	Overall
Text	9.7	7.6	7.1	8.0	11.7	7.6	8.6	14.7	8.9
PassPoints	13.7	14.1	13.3	15.3	13.7	14.1	13.9	19.3	14.3
PCCP	11.7	12.7	11.5	12.9	11.7	12.7	11.5	12.9	11.8
Faces	37.3	26.4	24.4	31.7	37.3	26.4	24.4	34.7	31.7
Things	29.5	24.7	26.2	32.4	31.2	24.7	26.2	43.2	28.0

Table 4: Interaction times and average total authentication times in seconds.