Patterns in the Wild: A Field Study of the Usability of Pattern and PIN-based Authentication on Mobile Devices

Emanuel von Zezschwitz[1], **Paul Dunphy**[2], **Alexander De Luca**[1],
[1]Media Informatics Group, University of Munich (LMU), Munich, Germany
[2]Culture Lab, Newcastle University, Newcastle upon Tyne, United Kingdom
{emanuel.von.zezschwitz, alexander.de.luca}@ifi.lmu.de, paul.dunphy@newcastle.ac.uk

ABSTRACT

Graphical password systems based upon the recall and reproduction of visual patterns (e.g. as seen on the Google Android platform) are assumed to have desirable usability and memorability properties. However, there are no empirical studies that explore whether this is actually the case on an everyday basis. In this paper, we present the results of a real world user study across 21 days that was conducted to gather such insight; we compared the performance of Android-like patterns to personal identification numbers (PIN), both on smartphones, in a field study. The quantitative results indicate that PIN outperforms the pattern lock when comparing input speed and error rates. However, the qualitative results suggest that users tend to accept this and are still in favor of the pattern lock to a certain extent. For instance, it was rated better in terms of ease-of-use, feedback and likeability. Most interestingly, even though the pattern lock does not provide any undo or cancel functionality, it was rated significantly better than PIN in terms of error recovery; this provides insight into the relationship between error prevention and error recovery in user authentication.

Author Keywords
PIN, Pattern, Authentication, Usability, Likeability

ACM Classification Keywords
H.5.2. Information Interfaces and Presentation: User Interfaces - Input devices and strategies, evaluation

General Terms
Human Factors; Performance; Security

INTRODUCTION

At the beginning of the mobile phone era, the sole purpose of devices was to provide call functionality to users. In recent years, they have evolved to become multi-purpose devices which carry at all times, private and sensitive information that require protection from unauthorized access [19]. The main protection mechanism that can be found on modern mobile devices is token based authentication (in the form of possession of the device itself) with a personal identification number (PIN). A number of observations can be made regarding PIN authentication: firstly, it is theoretically insecure as users typically only find a small range of PINs memorable [3], and they are vulnerable to attacks such as shoulder surfing [27]. Users themselves see the limitations of PINs and report a desire for higher security that requires minimal effort [5].

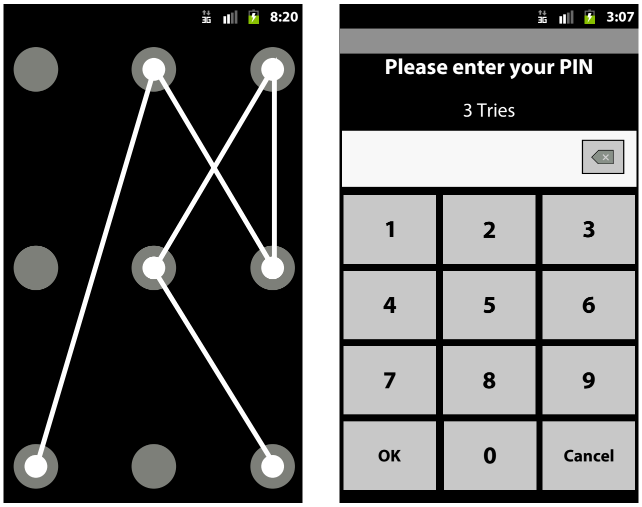


Figure 1. Screenshots of the prototype user interfaces of the pattern study (left) and the PIN study (right). The original interface language was German.

Graphical Passwords have received wide research attention as a candidate to provide usability and security in contexts including local authentication to mobile devices. The introduction of the Android operating system also brought the deployment of a drawmetric graphical password mechanism onto a platform whose market penetration is achieving staggering growth; in June 2012, Google announced a total number of over 400 million active Android devices and a growth of one million devices each day [6]. The *pattern lock* authentication mechanism is intended to prevent unauthorized access to Android devices, and requires users to use the touch screen functionality of the device to draw a memorable line pattern that connects dots displayed in a 3x3 configuration on-screen. The underlying credential is the sequence of dots hit whilst tracing the desired pattern; this system is arguably a constrained version of Draw-a-Secret [18] and other related systems (e.g. [16, 21, 32, 31]). Benefits of such an approach include that

MobileHCI '13, August 27 - 30 2013, Munich, Germany.
Copyright 2013 ACM 978-1-4503-2273-7/13/08$15.00.
http://dx.doi.org/10.1145/2493190.2493231

the system harnesses motor memory [29], the pictorial superiority effect [30], and has the potential to provide a larger theoretical keyspace than four-digit PINs. However, due to the lack of empirical work in this domain, it is unclear on what basis such systems provide benefits, especially to mobile device users, aside from providing a novel user experience.

In this paper, we present real world data on the performance of a PIN and a pattern lock on mobile devices. We assigned users either PINs or patterns and compared their authentication performance across 21 days; we provide a taxonomy of observed errors, and found that users made few errors with PINs and a relatively large number with patterns. The qualitative evaluation suggests that users are prepared to accept those errors, as the pattern mechanism was rated favorably in terms of ease-of-use, feedback, efficiency and memorability – which does not reflect what the quantitative performance measures would lead us to predict. The root of this contradiction appears to be the approach to *error recovery* adopted by the pattern-based authentication, which provides insight into a trade-off between error-prevention and error recovery when designing user authentication on mobile devices.

RELATED WORK

New authentication mechanisms are usually evaluated against two properties: *security* and *usability*. *Memorability* is a factor that influences both usability and security [1, 12]. For instance, security of an authentication system can be affected by memorability issues where users write down their authentication credentials, share them with others, or choose simple passwords or PINs [1]. Many user authentication systems exploit the so-called pictorial superiority effect [23] and/or the user's motor memory [29]; graphical passwords are one such example. De Angeli et al. [9] divided exemplar graphical password systems into locimetric, cognometric and drawmetric systems. Locimetric systems require the user to identify specific regions in an image; a very prominent example in this category is PassPoints by Wiedenbeck et al. [33]. Cognometric systems are based on the assumption that it is easy for people to recall and identify their own (password) images, that is, something known [24] amongst a larger set of decoy images. VIP by De Angeli et al. [8] and Déjà Vu by Dhamija et al. [14] are well-known examples in this category. Interestingly, in both categories, user-selected passwords are heavily dependent on the image content that is used which can lead to security problems [7, 26].

Drawmetric systems, the final category, are the main focus of this work. Like the work on locimetric and cognometric methods, they were originally designed to provide a more usable and memorable approach than PINs or passwords. Draw-a-Secret (DAS) by Jermyn et al. [18] can be considered the first of these systems. To authenticate, a user draws a shape within a 4x4 grid; the drawing does not need to be precisely repeated, however, the user must draw through the same grid cells in the same order. PassShapes by Weiss et al. [32] uses relative directions rather than coordinates. While the theoretical password space for DAS and other drawmetric systems is quite large, the user choice is predicted to reduce this space in practice [22, 25]. That is, when given the possibility, users choose simple shapes that are also easier to guess. Dunphy et al. [16] could show that background images partially solve this problem and, at the same time, increase the memorability of the shape.

Much work in the domain of usable and secure user authentication has focused upon making the input resistant to attacks like shoulder surfing [27], this is due to the ever increasing prevalence of mobile and ubiquitous computing contexts. Many of these systems rely on cleverly designed software [28, 34], additional hardware [2, 20] or hardware tokens in possession of the user [4, 13] to provide enhanced security. Unfortunately, this mostly correlates with reduced usability. To address shoulder surfing problems of drawmetric systems, Malek et al. [21] propose an approach in which a binary pressure code is applied when inputting the shape. That is, a stroke can be either drawn normally or with pressure, adding an invisible channel to the authentication process that is difficult to observe. In EyePassShapes [10], eye gaze is used as an invisible channel for user input; users perform the input with eye movement rather than with their hands. Finally, in [11], a biometric security layer was added to the pattern input, checking not only whether the correct pattern has been used but also how it has been input to identify whether the user is valid or not.

While short term and lab studies (e.g. [32]) have attested good memorability properties for drawmetric systems, none of the related work evaluated whether this holds true in a real world setting over a reasonable period of time. Thus, our work contributes to the field by presenting the results of such a study and discussing implications of the identified memorability, usability and user experience issues.

USER STUDY

We conducted a user study to collect user performance data away from a laboratory environment, and to gain insight into user perceptions of the usability and the likeability of the pattern approach; at the same time, we asked another set of participants to use a PIN system to provide a base for comparison.

Prototypes

The two prototypes, as shown in Figure 1, were designed based on typical implementations seen on mobile devices. In addition to *simple* patterns i.e. those permissible on Android instantiations of the pattern lock, our prototype also allowed entry of *complex* patterns, where dots could be skipped (see Figure 7) or visited several times. Thereby, we hoped to examine if the bigger password space is counter balanced by usability issues.The main difference between the pattern prototype and the standard Android implementation is that complex patterns are not possible on that platform, which reduces the size of the theoretical password space.

The prototypes had two modes: the training mode (allowing unlimited attempts), and afterwards, study mode. During study mode, 21 authentication sessions were allowed with a maximum of one per day. A session consisted of a maximum of three failed authentication attempts or one successful login.

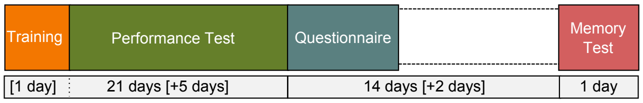


Figure 2. The time flow of the user study. Times in brackets are optional.

During use of the prototypes, log files were constantly created. For this, we stored every event happening on screen including touches, strokes, PINs, corrections, aborts and so on. Patterns were logged as a sequence of connected dots numbered from 1 to 9. The connection between two dots is called a stroke. The resulting log files were used to check for error rates, input speed etc. The study prototypes were both written for Android 2.1 or higher and distributed via a download link.

User Study Design

We wanted to compare PIN usage to pattern authentication; to do so we chose a between groups longitudinal design. Inside the pattern group, we randomly assigned complex and simple patterns. The task was to authenticate using the respective system once per day for 21 days. That is, the study went on for three consecutive weeks.

An important choice was on the complexity of the authentication tokens to assign to users. We decided to use four-digit PINs, the most common length found in real world systems. Having a theoretical password space of 10,000 different PINs, the password space for the patterns should be equal or higher than this number. The closest one can get to this using a 3x3 grid, is with five-stroke patterns with a theoretical password space of 32,768. This is based on the assumption that each of the other eight dots (nine minus the currently selected one) can be selected next and that returning to the previous dot is allowed. Once again, these numbers only represent the theoretical password space and as we randomly assigned passwords to our participants this password space remained at its full size. It has to be noted that in the real world, there are many factors significantly decreasing the password space [22, 25].

Finally, the data from users in this study was also used to identify biometric patterns in drawmetric systems, the results of which are published in [11].

Procedure

In a first step, we randomly assigned unique PINs or patterns (depending on the group) to each participant. By assigning patterns and PINs, we could control their complexity and assure the individuality of each token. On the first day of the study, the participants received an e-mail containing a download link for the respective prototype (pattern or PIN), an installation instruction, a manual, the unique PIN/pattern as well as the anonymous user ID used to connect the questionnaire answers to the log files. After installation, the participants had to input their ID (based on which the PIN/pattern was activated). In the next step, the study started with a training task, allowing the users to train their PIN or pattern without the results being logged. Whenever the participants felt ready, they could stop the training task and begin the actual study. After one day, the training task stopped automatically. In the following 21 days, the participants had to authenticate once per calendar day. Each time, a maximum of three attempts was allowed. To reduce the mental load for the users, an e-mail reminder was sent once every day. The reminder contained no personal data (e.g. authentication token). In case the participants still forgot the input, an extension of one day was granted. After five extensions, the respective person was removed from the study. That is, the study took a maximum of 27 days per participant.

Upon completion of this section of the study, participants were invited to the lab for a debriefing, during which the log files were extracted from their devices, and they were asked to fill in a questionnaire that collected qualitative data about how the participants perceived usability, performance, error-resistance, security and likeability of their assigned system. In some cases, participants were not able to appear in person (e.g. one participant was out of the country); in those cases, the approach was performed remotely with the experimenter providing detailed instruction on how to copy the log files and on how to access the questionnaire.

The final part of the study was not initially disclosed to the participants, indeed after the questionnaire participants assumed the study was complete, and we told participants that in the future we may have more questions for them regarding their study experience. After 14 days of non-use, we again arranged to meet each participant and asked them to recall their PIN/pattern using a printed version of the prototypes (see Figure 1); subsequently, participants were asked to rate the memorability of the respective system.

The meeting took place face-to-face and participants were allowed three attempts to correctly authenticate. We allowed up to a maximum of two additional days to pass to meet the participants. We decided to use a paper prototype as it was independent from any device specifications and could easily be taken along. Thus, the tests could be conducted anywhere whilst still maintaining the element of surprise with regard to the topic of the meeting. Figure 2 illustrates the different aspects and the time flow of the study.

Participants

Participants were recruited using mailing lists and social networks. As an incentive, in both groups, a gaming console was offered as a raffle prize; participants were invited to be present at the raffle. The only prerequisite to participate in the study was to own an Android mobile phone with Android 2.1 or higher. The two groups were not recruited at the same time but the PIN group started once the pattern group had finished.

For the pattern group, we managed to recruit 38 participants. 31 out of these finished the first part of the study including the final questionnaire. When performing the memorability test, we did not manage to get a hold of all remaining 31 participants. Thus, we had to remove another two participants leaving us with 29 valid data sets. The average age of the valid 29 participants was 26 years (19-36). Eleven participants were female, 18 male. 21% stated to use patterns on their smartphone to authenticate.

	Pattern	PIN
participants (training)	38	30
participants (performance test)	34	26
participants (memorability test)	29	24
average age	26 (19-36)	27 (21-42)
male/female	18/11	17/7

Table 1. Demographics and number of valid participants at the end of each stage of the study. Age and the male/female ratio are based on the final, valid participants.

The PIN group started with 30 participants, four of which had to be removed since they skipped more than 5 days. Out of the 26 that finished the first part including the questionnaire, one had to be removed due to an invalid dataset (only 18 days) and one participant missed the memory test. That is, the PIN study ended with 24 valid datasets. The average age of those 24 participants was 27 years (21-42). Seven of them were female, 17 male. 46% stated to use PIN on their smartphone to authenticate. Table 1 gives an overview of the statistical data we collected at the various stages of the study from both groups.

In the country where this study was conducted, there are no IRBs in place. They only exist in a few disciplines like medicine. Therefore, neither an IRB review nor an approval was required to conduct this study. However, it is required that studies with humans and their data abide to the German privacy laws. When designing and conducting the study, we took care that this held true for all parts of the process.

RESULTS

Quantitative data was collected via the built-in logging mechanisms. In addition, qualitative data was collected using a questionnaire and informal interviews.

Performance Test

To compare the performance of both approaches, input speed and failed authentications were analyzed.

Input Speed

In this section, we compare the average authentication times for both systems. Only correct attempts are included into the analysis. Since the pattern approach does not support to abort a running attempt, aborted attempts are among wrong attempts and thus often have very short input times. Therefore, including wrong attempts into the speed analysis would skew the results.

For the pattern scheme, time measurement started when the finger of the user touched the display and stopped as soon as the finger was lifted. Since the pattern scheme does not make use of an explicit confirmation step (e.g. press enter), this time was excluded for the analysis of the PIN approach to have comparable times. Therefore, the time measurement for PIN started when the user pressed the first numeric button and stopped when releasing the last button before the confirmation button was pressed.

A t-test revealed that people authenticated significantly faster using the PIN approach $t_{46.55} = -7.25$,$p < 0.001$, $r = 0.73$. On average, sessions using PIN tokens lasted 1501 ms (SE = 123 ms, Min = 844 ms, Max = 3141 ms) while sessions using the graphical approach lasted 3136 ms (SE = 189 ms, Min = 1257 ms, Max = 6184 ms). Taking into account undo operations, which were only supported by the PIN prototype, reveals that attempts where undo operations were used were significantly slower (M = 5720 ms, SE = 665 ms) than attempts without undo operations (M = 1314 ms, SE = 33 ms), $t_{24.12} = -6.62$, $p < 0.001$, $r = 0.80$. Overall, 20 corrected attempts (using backspace) and five cancelled attempts (using the cancel button) were logged within the PIN group.

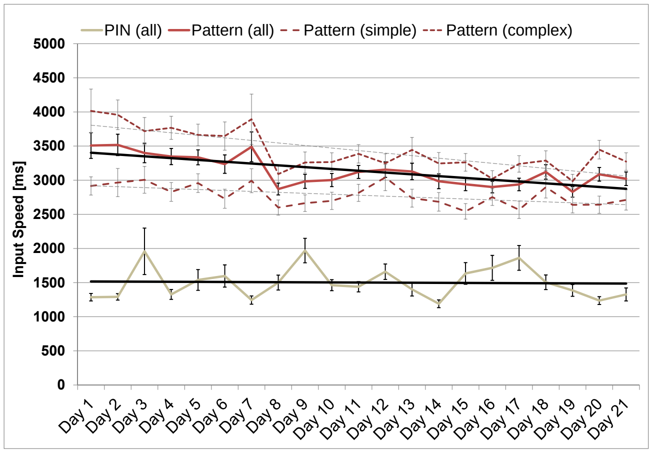


Figure 3. The average input times of the PIN/pattern group for all 21 test days. The black trend lines indicate a minor training effect for the pattern group within the first week of usage.

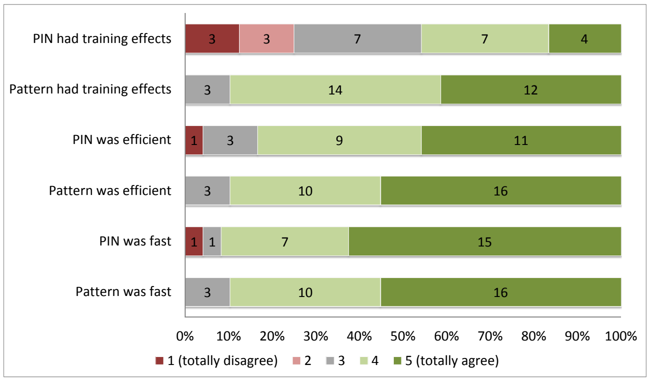


Figure 4. Qualitative results regarding the input speed of both systems.

Having a look at the users impressions taken from the questionnaire (see Figure 4), the different input speeds are not reflected by the users' opinion. Authentication speed was rated as fast or very fast by 21 (92%, Mdn = 5) PIN users and 26 (90%, Mdn = 5) pattern users, the difference is not significant according to a performed Wilcoxon rank-sum test, $W_s = 758.50$, $z = -0.50$, $p > 0.05$, $r = -0.07$. Neither the logged input times nor the users' ratings towards authentication speed were significantly influenced by the pattern complexity, $p > 0.05$. However complex patterns (e.g. skipped dots) led to slower average input times.

Another interesting finding of the questionnaire was that according to the Wilcoxon rank-sum test significantly more participants using the pattern scheme (90%, Mdn = 4) assessed their authentication speed to become faster over the three weeks period, $W_s = 472.50$, $z = -3.30$, $p < 0.001$, $r = -0.45$. Only one (4%, Mdn = 3) participant using the PIN approach stated the same.

To compare this finding with the quantitative log data, we conducted a one-way repeated measures ANOVA. The results do not support the users' impressions as there was no significant effect of the days on the input speed. This holds true for PIN, $F_{4.42,92.87} = 0.86$, $p > 0.05$, Greenhouse-Geisser corrected ($\epsilon = 0.22$), and for patterns, $F_{3.61,101.11} = 1.49$, $p > 0.05$, Greenhouse-Geisser corrected ($\epsilon = 0.18$). The complexity of the users' patterns had no significant effect as well ($p > 0.05$).

We conducted a second analysis to compare the input speed of the first day and the last day. It revealed that, compared to the first day (3508 ms, SE = 376 ms), pattern users performed perceptibly, but not significantly, faster (3021 ms, SE = 217 ms) at the end of the study, $t_{25} = 1.39$, $p > 0.05$. The correspondent times of the PIN group show only minor differences with 1286 ms (SE = 109 ms) on the first day and 1327 ms (SE = 190 ms) on the last day, $t_{23} = -0.20$, $p > 0.05$). This finding can partly explain the users' impressions and is visualized by the black trend lines seen in Figure 3. The complexity of the patterns had no significant influence ($p > 0.05$). Further analysis additionally revealed that the real-life usage of a certain token (PIN or pattern) had no significant influence on the input times of the study ($p > 0.05$).

Errors

The error rate is, like input speed, an important indicator to evaluate the usability of an authentication system. Beside the descriptive analysis, we focus on the distinction between slips and memory related errors. The security policy of the Android pattern system allows up to five failed attempts, before the system will be locked for 30 seconds. However, in terms of comparability, we decided to categorize errors according to the assumable consequences on a productive PIN-based system (e.g. an ATM). Three consecutive failed attempts are categorized as *critical failures* as they would usually lock such a system. One or two consecutive failed attempts, which are followed by a successful attempt, are categorized as *uncritical failures*.

We analyzed 504 (21 days $*$ 24 users) PIN-sessions and 609 (21 days $*$ 29 users) pattern-sessions. One session can consist of one, two or three attempts. With a total number of four (0.8%) failed sessions (two critical, two uncritical), the overall error rate of the PIN group is significantly smaller than the one of the pattern group, $t_{29.93} = -6.26$, $p < 0.001$, $r = 0.75$. Pattern users uncritically failed in 89 sessions (15%) and critically failed in ten (1.6%) sessions. The number of critical and uncritical failures was not significantly influenced by the pattern complexity ($p > 0.05$). The detailed analysis reveals that, in contrast to the number of critical failures, $t_{35.96} = -1.26$, $p > 0.05$, the number of uncritical failures is significantly influenced by the used token, $t_{29.59} = -7.10$, $p < 0.05$, $r = 0.79$. Figure 6 shows the daily error rate of both approaches.

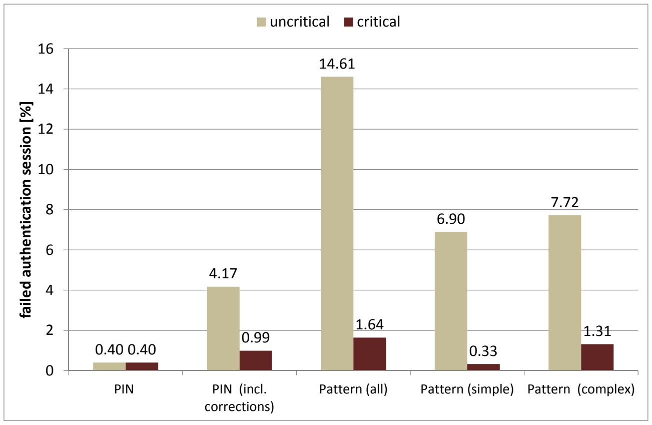


Figure 5. Authentication sessions including failed or corrected & failed (PIN only) attempts.

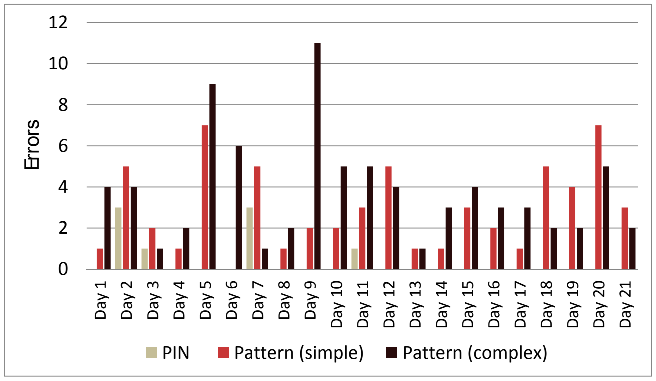


Figure 6. The error rate of PIN, pattern (simple) and pattern (complex) for all 21 test days.

It should not be ignored that, in contrast to PIN, the pattern approach does not support undo operations (e.g. cancel, backspace). This is why aborted attempts are among failures in the pattern group. To adjust this factor, we performed a second analysis, counting cancelled and corrected PIN attempts as failures. Including those attempts, the PIN group performed five (0.1%) sessions with a critical failure and 21 (4.2%) sessions with an uncritical failure. This is still significantly better, $t_{51} = 2.81$, $p < 0.05$, $r = 0.37$. Figure 5 illustrates the error rates of both groups.

To examine the reasons for the large number of failures in the pattern group, we developed a novel taxonomy for the categorization of errors. Since the scheme is based on theoretical assumptions, we asked four randomly chosen participants to confirm the categorization of their errors. According to the mapping of a standard PIN-pad (figure 1, right), we illustrate the errors based on the sample pattern "1 2 3 6 4 7" as depicted in figure 7, left.

1. **Slips are based on**...

   (a) aborting a pattern ("1 2 3 6").

   (b) a correct pattern which is distributed over two or three attempts ("1 2 3" for the first attempt, "6 4 7" for the second attempt).

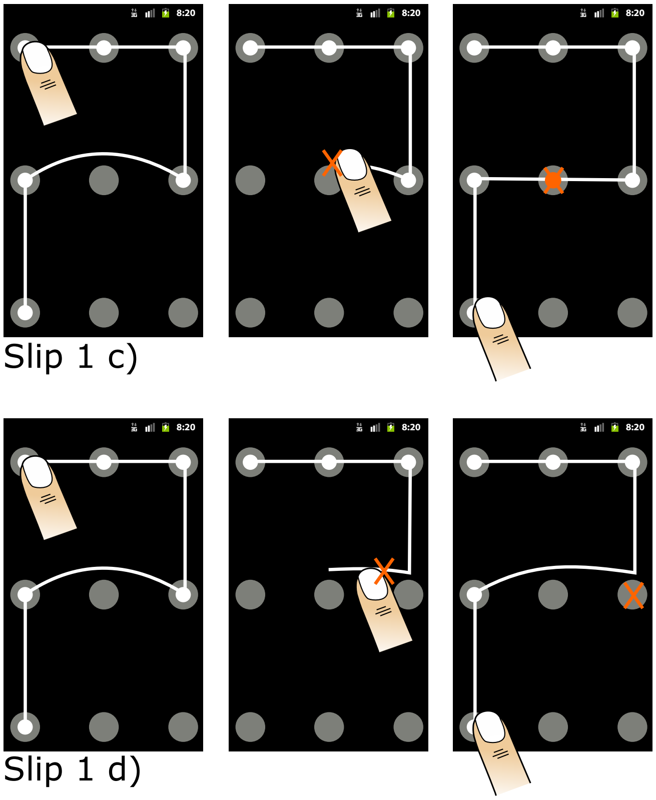


Figure 7. Two common slips when using the pattern system. Top: The user tries to skip a dot but accidentally touches it, creating a wrong input. This has been defined as slip 1 c) in our categorization. Bottom: The user misses out on one dot creating a too short pattern (1 d) in the categorization).

(c) inserting additional strokes ("1 2 3 6 5 4 7"). See figure 7 (top) for an example.

(d) missing strokes ("1 2 3 5 4 7"). See figure 7 (bottom).

(e) only one wrong stroke which connects a wrong point near the correct point ("1 2 3 6 4 8").

2. **Memory errors are based on**. . .

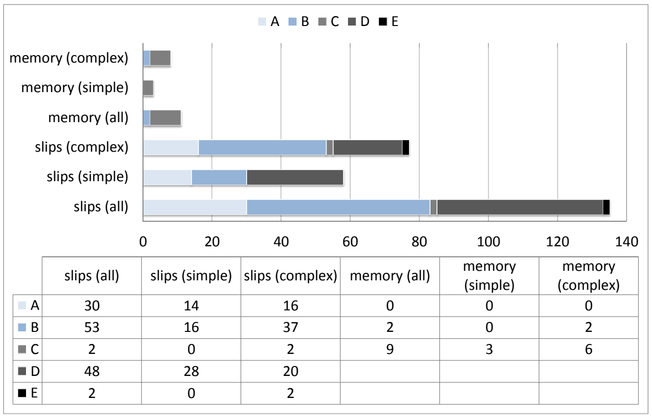
(a) a repetition of the same wrong pattern ("4 5 6 9 8 5" for two or three attempts).

(b) one or more wrong strokes which are not direct neighbors of the correct stroke ("1 2 5 9 4 3").

(c) wrong strokes that are mirrored versions of the correct strokes ("3 2 1 4 6 9").

Using the described scheme, we categorized 135 (93%) slips and 11 (7.5%) memory errors in the pattern group. 58 (43%) slips and three (27%) memory errors were based on simple patterns (see figure 8). Whenever an error fitted in both categories (slips and memory), we counted it as a memory error. That is, because memory errors often are followed by user behavior we would usually categorize as slips. For example, there have been attempts which were aborted after a mirrored stroke.

As seen in figure 8, 53 (40%) slips are patterns that were torn apart and thus were logged as separate attempts. 16 (30%) of these were based on simple patterns. Those errors occur when an the finger of the user is briefly lifted from the display such that the login is interrupted, and the participant does not become aware that the login has restarted. Since the system has no explicit confirmation mechanism, each such interruption is logged as a failed authentication attempt. Other frequently committed slips are missing strokes (36%), 28 (58%) of these based on simple patterns. 30 (22%) errors are based on aborted attempts, 14 (47%) of these are simple. Amongst aborted attempts are wrong and correct stroke sequences. While the abort of a correct pattern seems to be based on unintended interruptions (slips), the abort of wrong patterns is based on recognized errors. Therefore, the abort of a wrong pattern seems to be intended and is rather a consequence of a previous failure than an error itself. One common error which lead to the abort of many attempts, was accidentally touching a point which actually should have been skipped (Figure 7). All but one critical failure has been the consequence of slips.



	slips (all)	slips (simple)	slips (complex)	memory (all)	memory (simple)	memory (complex)
■ A	30	14	16	0	0	0
■ B	53	16	37	2	0	2
■ C	2	0	2	9	3	6
■ D	48	28	20			
■ E	2	0	2			

Figure 8. The distribution of slips and memory errors within the pattern group. The letters stand for the respective category as introduced in the error analysis.

In the PIN group, six failed attempts were logged. One critical error was based on a memory problem. The user transposed the digits of the correct PIN. Consequently, it is related to category 2b). The second critical failure was based on a hardware problem and thus does not fit in any category. Both uncritical failures are based on missing digits and are therefore related to category 1d).

Questionnaire

In a questionnaire we probed aspects of the participants' experience with both mechanisms, in categories of usability, likeability and memorability.

Usability

Analyzing the relevant answers of the questionnaire (see Figure 9) reveals that pattern users were not irritated by the number of failed attempts. 27 (93%, Mdn = 5) pattern users and 21 PIN users (88%, Mdn = 4) stated that the system was easy to use, $W_s = 595.50$, $z = -1.05$,$p > 0.05$, $r = -0.14$. Significantly more pattern users (90%, Mdn = 5) stated that errors could be quickly recovered from, $W_s = 507.00$, $z = -2.73$, $p < 0.05$, $r = -0.37$. Only 54% (Mdn = 4) of the PIN users stated the same. In addition 97% (Mdn = 5) of the pat-

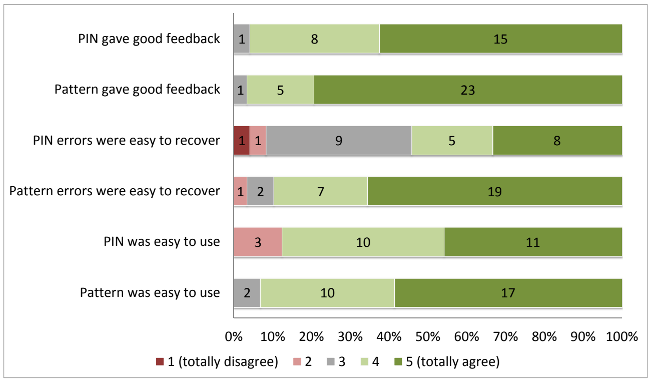


Figure 9. Qualitative results regarding the error rate of both systems.

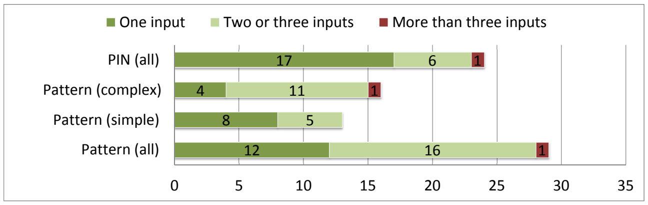


Figure 10. Self-reported number of inputs the participants needed to memorize their token.

tern group (96% of the PIN group, Mdn = 5) rated the feedback of the application very comprehensible, $W_s = 507.00$, $p > 0.05$, and 79% (Mdn = 5) of the pattern users (54% of the PIN users, Mdn = 4) acclaimed that error messages clearly defined the current problem. Even if the perception does not differ significantly ($p > 0.05$) this is an unexpected rating of the pattern group, considering that 39% of all slips happened because the feedback of the system was ignored or missed (category 1b)).

Memorability

While, according to the qualitative error analysis, only 11 memory errors could be derived from the log data, this section describes the evaluation of the generic memorability of both approaches. The analysis is based on the questionnaire and the recall test which was conducted two weeks after finishing the performance test.

According to the users' statements (see Figure 10), PIN users needed significantly less time to memorize their token, $W_s = 550.50$, $z = -1.99$, $p < 0.05$,$r = -0.27$. 41% of the pattern users (67% simple) and 70% of the PIN users could memorize their token after the first input. 55% (31% simple) of the pattern group participants and 25% of the PIN group needed two or three inputs. In both groups, there was only one participant who stated that the memorization of the token took more than 3 inputs. This data supports the results of the qualitative error analysis which revealed that most of the errors in both groups were slips. However, simple patterns seem to be easier to remember than their more complex counterparts.

All participants of the pattern group, who ranked the graphical approach, said that it was easy to memorize. None of them rated PIN to be easier than patterns. Three participants explicitly stated they would visually memorize PINs, too. However, remembering the pattern is not a guarantee to succeed. In one case, the participant could recall his pattern, but did not remember its starting point. As a consequence, the authentication failed. According to the recall test, patterns seem to be easy to memorize, but there is no indication for the assumption that patterns are easier to recall than PINs.

Likeability

The perceived performance indicates that PINs and patterns work equally well. According to the participants' rating, patterns are fast, easy to use and, in contrast to PINs, errors are recovered from quickly. However, the findings of the log data seem to contradict the users' impressions as pattern users authenticated significantly slower and committed significantly more errors than PIN users. This section shows that one reason for the apparent contrasts of the logged data and the users' perceptions is likeability. Thereby, ratings of specific aspects do not necessarily correspond to the overall likeability ratings.

As seen in Figure 11, 59% (Mdn = 4) of the pattern users and 71% of the PIN users (Mdn = 4) liked interacting with the user interface. The Wilcoxon rank-sum test reveals that the difference is not significant, $W_s = 630.50$, $z = -0.33$, $p > 0.05$, $r = -0.05$. The rating is not influenced by the pattern-complexity, but might be influenced by usability issues. This is supported by the reports of some participants who stated that some dots were very distant and thus hard to reach. The layout of the graphical user interface was rated fairly similar for both prototypes. 62% of the pattern users and 63% of the PIN users liked the appearance of the used GUI, $p > 0.05$.

Taking both ratings into account, one could assume that likeability ratings are in the benefit of the PIN approach. Anyhow, when we asked the users how they felt using the prototype, 86% of the pattern users (Mdn = 5) and only 75% of the PIN users (Mdn = 4) stated that it felt good. Even if the difference is not significant, $W_s = 589.00$, $z = -1.15$, $p > 0.05$, $r = -0.16$, this is an unexpected result as it does not reflect the above described ratings. In addition, the overall likeability is confirmed by fact that 90% of the pattern group and 83% of the PIN group were pleased with the tested system, $p > 0.05$.

Memory Test

According to the spontaneous recall test, the memorability of both tokens seems to be fairly equal. In the PIN group, 22 participants (92%) remembered their token. 19 (86%) among them needed only one attempt. In the pattern group, 26 participants (90%) could recall their credential. 23 (89%) among them gave the correct answer on the first attempt. A complex pattern was assigned to two of the three pattern users who could not recall their token. Asked for their strategies to memorize, 16 (67%) participants of the PIN group stated that they memorized their PIN on a motor or visual basis. That is, they recalled the pattern, which evolved from connecting the intended buttons with imaginary lines. Others remembered their PIN by associating previous knowledge or simply "learned it by heart".

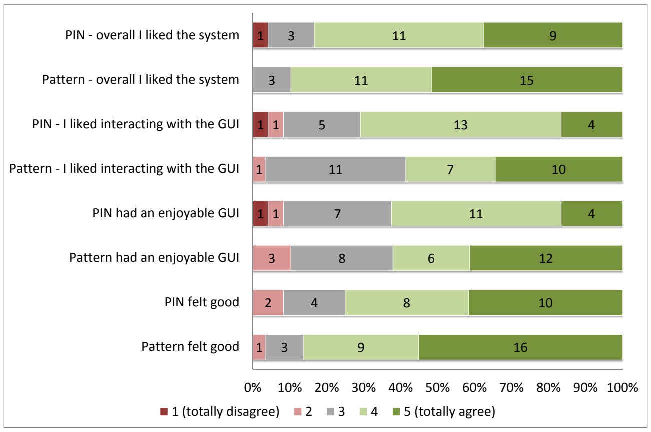


Figure 11. The participants' perception of patterns and PINs.

DISCUSSION

We reported results of daily usage of PIN and pattern-based authentication across 21 days, and the results of a spontaneous memory test administered two weeks later. Our study was designed to capture the time period after the assignment of new credentials which can be a difficult time for committing new credentials to memory. During that time, we observed that the system-type significantly influenced the success rate and the input speed of participants. In terms of input speed we observed that the mean daily login duration showed improvement across the 21 days, with the mean duration of a successful login decreasing from 3.5 seconds initially, to 3 seconds at the end of the study. Such an improvement was not observed in the PIN group, where the mean successful login duration remained similar at the start and end of the study at approximately 1.25 seconds. Overall, users of the pattern system needed more than twice as much time as PIN users to achieve a successful login. Despite this, the mean login duration of pattern (M = 3136 ms) was still fast in comparison to other graphical password schemes [17, 15].

In terms of success rates, pattern users made significantly more errors than users of PIN. Analyzing patterns based upon whether they were simple patterns showed that users of patterns that required more complicated interactions did not make significantly more errors than those using simple patterns. This is encouraging as it suggests that allowing more complex (and thus more secure) drawings in this study did not impede the memorability or usability of those patterns. We devised a novel taxonomy of the errors users would be likely to make when using pattern-based authentication. Data from our study suggests that slips involving the user lifting their finger in the middle of a stroke were a significant cause of error, along with hitting a dot closely positioned to the intended dot. It is possible that software intervention could alleviate the effects of these errors, although these could have been exacerbated due to the area across which the dots were distributed.

Despite the errors observed in user interactions with the pattern-based approach, it appeared that there remained a strong preference for the pattern system. In the questionnaire we distributed that elicited participant perceptions regarding their assigned system, on the surface, pattern scored better in terms of: ease of use, quality of feedback, and various dimensions of likeability. One possible source of such positive reaction (when discounting a lab-effect) is a significant result when comparing scores regarding the *error recovery* aspect of usability. Pattern users rated the error recovery features of pattern significantly higher than PIN users rated those of PIN. Error recovery is likely to have a number of dimensions in the context of user authentication:

- Corrective Interactions - This could involve moving a cursor to delete a particular character etc., press backspace, navigate screens to undo a particular selection
- Reset Interactions - The number of interactions required to put the interface in a state to enter the credentials afresh. This could involve interaction with modal dialogs, re-entry of usernames
- Loading Time - The user may have to wait for credentials to be checked on a server, be subject to a random delay for too many guesses.
- Feedback - Can the user understand what aspect of entry was incorrect in order to correct their behavior?

Using security mechanisms in the wild places very different constraints upon the user, and login errors are more likely due to the multitude of distractions in the environment. Typically, in studies of user authentication, the enumeration of errors is the key focus, however pattern authentication appears to represent a trade-off between designing to prevent errors and to recover from errors. In the case of our prototype and the Android pattern interface, both provide no error prevention and focus upon a speedy error recovery. A comparison of pattern and PIN in this regard is seen in table 2. It appears that pattern-based authentication implicitly provides good feedback to users regarding mistakes made during the login procedure through the visual appearance of the pattern onscreen. In the case of PIN, each digit is obscured through the digits themselves being masked through an onscreen dot, which could make login errors more frustrating as the source of the error is unknown. Such design decisions have implications for observation attacks, too.

Where PIN users made use of corrective or reset interactions (i.e. undo or clear), the recorded mean login durations were significantly slower than when such functions were not used (5720 ms vs. 1314 ms). Correcting errors resulted in login durations over four times the length of a typical login. In the context of Android patterns and our own prototype, the need to correct mistakes, succumbing to distractions etc. immediately caused login failures and a resetting of the interface ready for re-entry of credentials. Prioritizing a speedy recovery from errors ahead of avoiding errors, raises interesting questions about the usefulness of success rates in field studies. Informal observations from an earlier field study [15] of

	Pattern	PIN
Corrective Interaction	None	Backspace
Reset Interactions	None	Clear
Loading Time	None	None
Feedback	visible pattern	yes/no feedback

Table 2. Comparison of recovery features of pattern and PIN.

recognition-based graphical passwords suggests that participants were most likely to force a failed login when making a mistake rather than navigate to the 'start again' option because it was more convenient which impacted success rates. In particular, this forces reflection upon traditional *three strikes* lockout usually attached to authentication failures. This suggests that designers of user authentication mechanisms should carefully consider the process of recovering from errors, and view with more scrutiny the errors that users are observed to make in everyday usage.

STUDY LIMITATIONS

User studies in the wild can create interesting results but, at the same time, it is difficult to control all the variables that may impact user performance. One of the main goals of the study was to find long term effects of using PIN and pattern-based authentication. The question remains whether three weeks was a sufficient timespan for this purpose. However, we had to choose this time frame to best manage the likelihood of high dropout rates. In addition, and related to this, we did not test for the influence of very long breaks on PIN and shape authentication performance. Rarely used authentication credentials tend to produce the most problems for users when authenticating to a system. The memorability test towards the end of the study gave interesting insight into the memorability of both systems under regular usage, but it would be also very interesting to find out how the systems perform after longer non-use, say, one year or more.

Another thing we could and did not test for is interference effects. Even though the participants did not use their own PINs and patterns, we cannot say for sure whether multiple PINs or patterns would have influenced the performance. An important point here is that users already have and use multiple PINs for bank cards, credit cards, their mobile devices etcetera while it is very likely that they had a maximum of one other pattern to remember. The biggest effect that we were not able to control is the fact that people in general are highly trained in using PIN authentication, and unfamiliar with drawmetric graphical passwords; this could have influenced perceptions of performance and likeability. Future study can help us determine whether this effect holds true in a different user samples. Also, although the sample size of both study groups could be larger, it has proved sufficient to generate statistically significant results.

In addition, implementation issues could have influenced the performance. As opposed to the native authentication mechanisms on a mobile device, we could not make sure that our prototypes ran independently from other processes. Finally, as our system allowed a wider range of patterns than Google's pattern authentication, the results might not thoroughly be generalizable to this specific authentication scheme, even though complex patterns had no significant impact.

CONCLUSION AND FUTURE WORK

In this paper, we presented the results of a field study, in which we evaluated performance, usability and likeability aspects of Android-like graphical passwords in the wild in comparison with PIN. In addition, we presented a taxonomy that helps to get further insights into the origins of logged errors. By evaluating Android-conform patterns, we gathered generalizable results, while the evaluation of their more complex counterparts provided insights into the effects of allowing a non-restricted password space.

The study revealed that one main difference of both approaches is the concept of error recovery. While the PIN prototype allows for recovering from errors using undo operations, the pattern users were forced to submit every attempt without corrections. This is the same approach that is found in current real world implementations. The results show that input speed and success rate were influenced by the concept of error recovery. While the average input speed of PIN users was significantly faster, we could show that using undo operations had a significant effect on the input speed as well. Thus, the average time of authentications, where such operations were used, were significantly slower, even compared to sessions using the pattern approach.

Using our taxonomy revealed that most of the errors of the pattern group were based on slips and thus, one could assume that a lot of these errors could have been avoided using undo operations. However as mentioned above, undo operations were not supported by the pattern prototype. Even if they could have been avoided, based on our qualitative findings, it is doubtful that they would have been avoided very often. The findings revealed that significantly more participants of the pattern group stated that recovering from errors was fast or very fast. In addition, the pattern prototype users were not irritated by the number of failed attempts as likeability ratings benefited the pattern prototype as well. This leads to the conclusion that fast error recovery is more important for the users than error avoidance and questions the benefit of undo operations for such an approach.

There are a couple of open points that we consider worth investigating in the future. Firstly, since our study revealed first interesting insights into this matter, we would like to further examine the different kinds of error recovery and its implications. Related to this, real world logging of failed authentication attempts using the operating systems' native implementations and the users' own patterns and PINs would be the optimal solution to do this. However, this has to be done with respect to the users' privacy. Finally, as mentioned before, we do not know anything about interference effects between multiple different Android (or Android-like) patterns. Only after examining this, we can say for sure whether the memorability properties of such a drawmetric system are as good as commonly assumed.

ACKNOWLEDGMENTS

This work was partially funded by a Google Research Award. Special thanks go to Christian Lindner for his valuable help with parts of the user study.

REFERENCES

1. Adams, A., and Sasse, M. A. Users are not the enemy. *Commun. ACM 42*, 12 (1999), 40–46.
2. Bianchi, A., Oakley, I., and Kwon, D. S. The secure haptic keypad: a tactile password system. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, ACM (New York, NY, USA, 2010), 1089–1092.
3. Bonneau, J., Preibusch, S., and Anderson, R. A birthday present every eleven wallets? the security of customer-chosen banking PINs. In *16th International Conference on Financial Cryptography.*, Springer-Verlag (Heidelberg, Germany, 2012).
4. Chong, M., and Marsden, G. Exploring the use of discrete gestures for authentication. In *Human Computer Interaction - INTERACT 2009*, vol. 5727, Springer Berlin Heidelberg (2009), 205 – 213.
5. Clarke, N., Furnell, S., Rodwell, P., and Reynolds, P. Acceptance of subscriber authentication methods for mobile telephony devices. *Computers & Security 21*, 3 (2002), 220–228.
6. Cutler, K.-M. Android reaches 400 million device activations adds 1 million per day. Website, June 2012. Available online at http://techcrunch.com/2012/06/27/android-reaches-400-million-devices-activations-1-million-per-day; visited on August 6th 2012.
7. Davis, D., Monrose, F., and Reiter, M. K. On user choice in graphical password schemes. In *SSYM'04: Proceedings of the 13th conference on USENIX Security Symposium*, USENIX Association (Berkeley, CA, USA, 2004), 11–11.
8. De Angeli, A., Coutts, M., Coventry, L., Johnson, G. I., Cameron, D., and Fischer, M. H. Vip: a visual approach to user authentication. In *AVI '02: Proceedings of the Working Conference on Advanced Visual Interfaces*, ACM (New York, NY, USA, 2002), 316–323.
9. De Angeli, A., Coventry, L., Johnson, G., and Renaud, K. Is a picture really worth a thousand words? exploring the feasibility of graphical authentication systems. *Int. J. Hum.-Comput. Stud. 63*, 1-2 (2005), 128–152.
10. De Luca, A., Denzel, M., and Hussmann, H. Look into my eyes!: can you guess my password? In *SOUPS '09: Proceedings of the 5th Symposium on Usable Privacy and Security*, ACM (New York, NY, USA, 2009), 1–12.
11. De Luca, A., Hang, A., Brudy, F., Lindner, C., and Hussmann, H. Touch me once and i know it's you! implicit authentication based on touch screen patterns. In *Proceedings of the 2012 annual conference on Human factors in computing systems*, CHI '12, ACM (New York, NY, USA, 2012).
12. De Luca, A., Langheinrich, M., and Hussmann, H. Towards understanding atm security: a field study of real world atm use. In *SOUPS '10: Proceedings of the Sixth Symposium on Usable Privacy and Security*, ACM (New York, NY, USA, 2010), 1–10.
13. De Luca, A., von Zezschwitz, E., and Hussmann, H. Vibrapass: secure authentication based on shared lies. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, ACM (New York, NY, USA, 2009), 913–916.
14. Dhamija, R., and Perrig, A. Déjà vu: a user study using images for authentication. In *SSYM'00: Proceedings of the 9th conference on USENIX Security Symposium*, USENIX Association (Berkeley, CA, USA, 2000), 4–4.
15. Dunphy, P., Heiner, A. P., and Asokan, N. A closer look at recognition-based graphical passwords on mobile devices. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, ACM (New York, NY, USA, 2010), 1–12.
16. Dunphy, P., and Yan, J. Do background images improve "draw a secret" graphical passwords? In *CCS '07: Proceedings of the 14th ACM conference on Computer and communications security*, ACM (New York, NY, USA, 2007), 36–47.
17. Hayashi, E., Dhamija, R., Christin, N., and Perrig, A. Use your illusion: secure authentication usable anywhere. In *Proceedings of the 4th symposium on Usable privacy and security*, SOUPS '08, ACM (New York, NY, USA, 2008), 35–45.
18. Jermyn, I., Mayer, A., Monrose, F., Reiter, M. K., and Rubin, A. D. The design and analysis of graphical passwords. In *SSYM'99: Proceedings of the 8th conference on USENIX Security Symposium*, USENIX Association (Berkeley, CA, USA, 1999), 1–1.
19. Karlson, A. K., Brush, A. B., and Schechter, S. Can i borrow your phone?: understanding concerns when sharing mobile phones. In *Proceedings of the 27th international conference on Human factors in computing systems*, CHI '09, ACM (New York, NY, USA, 2009), 1647–1650.
20. Kim, D., Dunphy, P., Briggs, P., Hook, J., Nicholson, J., Nicholson, J., and Olivier, P. Multi-touch authentication on tabletops. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, ACM (New York, NY, USA, 2010), 1093–1102.
21. Malek, B., Orozco, M., and El Saddik, A. Novel shoulder-surfing resistant haptic-based graphical password. In *EuroHaptics 2006* (July 2006).
22. Nali, D., and Thorpe, J. Analyzing user choice in graphical passwords. Tech. rep., School of Computer Science, Carleton University, 2004.
23. Nelson, D. L., Reed, V. S., and Walling, J. R. Pictorial superiority effect. *Journal of Experimental Psychology: Human Learning and Memory 2*, 5 (Sept. 1976), 523–528.
24. Norman, D. *The Design of Everyday Things*. Perseus Books, Aug. 2002.
25. Oorschot, P. C. v., and Thorpe, J. On predictive models and user-drawn graphical passwords. *ACM Trans. Inf. Syst. Secur. 10* (January 2008), 5:1–5:33.
26. Renaud, K., and De Angeli, A. Visual passwords: cure-all or snake-oil? *Commun. ACM 52*, 12 (Dec. 2009), 135–140.
27. Rogers, J. Please enter your 4-digit pin. *Financial Services Technology, U.S. Edition Issue 4* (Mar. 2007).
28. Roth, V., Richter, K., and Freidinger, R. A pin-entry method resilient against shoulder surfing. In *CCS '04: Proceedings of the 11th ACM conference on Computer and communications security*, ACM (New York, NY, USA, 2004), 236–245.
29. Shadmehr, R., and Brashers-krug, T. Functional stages in the formation of human long-term motor memory. *The Journal of Neuroscience 17* (1997), 409–419.
30. Standing, L. Learning 10,000 pictures. *The Quarterly Journal of Experimental Psychology 25* (1973), 203–222.
31. Tao, H., and Adams, C. Pass-go: A proposal to improve the usability of graphical passwords. *International Journal of Network Security 7*, 2 (2008), 273–292.
32. Weiss, R., and De Luca, A. Passshapes - utilizing stroke based authentication to increase password memorability. In *NordiCHI 2008: Proceedings of the 5th Nordic Conference on Human-Computer Interaction*, ACM (New York, NY, USA, 2008), 383–392.
33. Wiedenbeck, S., Waters, J., Birget, J., Brodskiy, A., and Memon, N. PassPoints: design and longitudinal evaluation of a graphical password system. *International Journal of Human-Computer Studies 63*, 1-2 (July 2005), 102–127.
34. Wiedenbeck, S., Waters, J., Sobrado, L., and Birget, J.-C. Design and evaluation of a shoulder-surfing resistant graphical password scheme. In *AVI '06: Proceedings of the working conference on Advanced visual interfaces*, ACM (New York, NY, USA, 2006), 177–184.