Risks and Potentials of Graphical and Gesture-based Authentication for Touchscreen Mobile Devices

Balancing Usability and Security through User-centered Analysis and Design

Dissertation

an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

vorgelegt von
Diplom-Medieninformatiker

Emanuel von Zezschwitz
geboren am 15. Januar 1984 in München

München, den 24. August 2016

Erstgutachter:	Prof. Dr. Heinrich Hußmann
Zweitgutachter:	Prof. Dr. Matthew Smith
Drittgutachter:	Prof. Andrea Bianchi, Ph.D

Tag der mündlichen Prüfung: 07.11.2016

ABSTRACT

While a few years ago, mobile phones were mainly used for making phone calls and texting short messages, the functionality of mobile devices has massively grown. We are surfing the web, sending emails and we are checking our bank accounts on the go. As a consequence, these internet-enabled devices store a lot of potentially sensitive data and require enhanced protection. We argue that authentication often represents the only countermeasure to protect mobile devices from unwanted access.

Knowledge-based concepts (e.g., PIN) are the most used authentication schemes on mobile devices. They serve as the main protection barrier for many users and represent the fallback solution whenever alternative mechanisms fail (e.g., fingerprint recognition). This thesis focuses on the risks and potentials of gesture-based authentication concepts that particularly exploit the touch feature of mobile devices. The contribution of our work is threefold. Firstly, the problem space of mobile authentication is explored. Secondly, the design space is systematically evaluated utilizing interactive prototypes. Finally, we provide generalized insights into the impact of specific design factors and present recommendations for the design and the evaluation of graphical gesture-based authentication mechanisms.

The problem space exploration is based on four research projects that reveal important real-world issues of gesture-based authentication on mobile devices. The first part focuses on authentication behavior in the wild and shows that the mobile context makes great demands on the usability of authentication concepts. The second part explores usability features of established concepts and indicates that gesture-based approaches have several benefits in the mobile context. The third part focuses on observability and presents a prediction model for the vulnerability of a given grid-based gesture. Finally, the fourth part investigates the predictability of user-selected gesture-based secrets.

The design space exploration is based on a design-oriented research approach and presents several practical solutions to existing real-world problems. The novel authentication mechanisms are implemented into working prototypes and evaluated in the lab and the field. In the first part, we discuss smudge attacks and present alternative authentication concepts that are significantly more secure against such attacks. The second part focuses on observation attacks. We illustrate how relative touch gestures can support eyes-free authentication and how they can be utilized to make traditional PIN-entry secure against observation attacks. The third part addresses the problem of predictable gesture choice and presents two concepts which nudge users to select a more diverse set of gestures.

Finally, the results of the basic research and the design-oriented applied research are combined to discuss the interconnection of design space and problem space. We contribute by outlining crucial requirements for mobile authentication mechanisms and present empirically proven objectives for future designs. In addition, we illustrate a systematic goal-oriented development process and provide recommendations for the evaluation of authentication on mobile devices.

ZUSAMMENFASSUNG

Während Mobiltelefone vor einigen Jahren noch fast ausschließlich zum Telefonieren und zum SMS schreiben genutzt wurden, sind die Anwendungsmöglichkeiten von Mobilgeräten in den letzten Jahren erheblich gewachsen. Wir surfen unterwegs im Netz, senden E-Mails und überprüfen Bankkonten. In der Folge speichern moderne internetfähigen Mobilgeräte eine Vielfalt potenziell sensibler Daten und erfordern einen erhöhten Schutz. In diesem Zusammenhang stellen Authentifizierungsmethoden häufig die einzige Möglichkeit dar, um Mobilgeräte vor ungewolltem Zugriff zu schützen.

Wissensbasierte Konzepte (bspw. PIN) sind die meistgenutzten Authentifizierungssysteme auf Mobilgeräten. Sie stellen für viele Nutzer den einzigen Schutzmechanismus dar und dienen als Ersatzlösung, wenn alternative Systeme (bspw. Fingerabdruckerkennung) versagen. Diese Dissertation befasst sich mit den Risiken und Potenzialen gestenbasierter Konzepte, welche insbesondere die Touch-Funktion moderner Mobilgeräte ausschöpfen. Der wissenschaftliche Beitrag dieser Arbeit ist vielschichtig. Zum einen wird der Problemraum mobiler Authentifizierung erforscht. Zum anderen wird der Gestaltungsraum anhand interaktiver Prototypen systematisch evaluiert. Schließlich stellen wir generelle Einsichten bezüglich des Einflusses bestimmter Gestaltungsaspekte dar und geben Empfehlungen für die Gestaltung und Bewertung grafischer gestenbasierter Authentifizierungsmechanismen.

Die Untersuchung des Problemraums basiert auf vier Forschungsprojekten, welche praktische Probleme gestenbasierter Authentifizierung offenbaren. Der erste Teil befasst sich mit dem Authentifizierungsverhalten im Alltag und zeigt, dass der mobile Kontext hohe Ansprüche an die Benutzerfreundlichkeit eines Authentifizierungssystems stellt. Der zweite Teil beschäftigt sich mit der Benutzerfreundlichkeit etablierter Methoden und deutet darauf hin, dass gestenbasierte Konzepte vor allem im mobilen Bereich besondere Vorzüge bieten. Im dritten Teil untersuchen wir die Beobachtbarkeit gestenbasierter Eingabe und präsentieren ein Vorhersagemodell, welches die Angreifbarkeit einer gegebenen rasterbasierten Geste abschätzt. Schließlich beschäftigen wir uns mit der Erratbarkeit nutzerselektierter Gesten.

Die Untersuchung des Gestaltungsraums basiert auf einem gestaltungsorientierten Forschungsansatz, welcher zu mehreren praxisgerechte Lösungen führt. Die neuartigen Authentifizierungskonzepte werden als interaktive Prototypen umgesetzt und in Labor- und Feldversuchen evaluiert. Im ersten Teil diskutieren wir Fettfingerattacken ("smudge attacks") und präsentieren alternative Authentifizierungskonzepte, welche effektiv vor diesen Angriffen schützen. Der zweite Teil beschäftigt sich mit Angriffen durch Beobachtung und verdeutlicht wie relative Gesten dazu genutzt werden können, um blickfreie Authentifizierung zu gewährleisten oder um PIN-Eingaben vor Beobachtung zu schützen. Der dritte Teil beschäftigt sich mit dem Problem der vorhersehbaren Gestenwahl und präsentiert zwei Konzepte, welche Nutzer dazu bringen verschiedenartige Gesten zu wählen.

Die Ergebnisse der Grundlagenforschung und der gestaltungsorientierten angewandten Forschung werden schließlich verknüpft, um die Verzahnung von Gestaltungsraum und Problemraum zu diskutieren. Wir präsentieren wichtige Anforderungen für mobile Authentifizierungsmechanismen und erläutern empirisch nachgewiesene Zielvorgaben für zukünftige Konzepte. Zusätzlich zeigen wir einen zielgerichteten Entwicklungsprozess auf, welcher bei der Entwicklung neuartiger Konzepte helfen wird und geben Empfehlungen für die Evaluation mobiler Authentifizierungsmethoden.

Statement of Collaboration

This thesis presents the results of the research I carried out between January 2012 and July 2016. However, the wide range of empirical results would not have been possible without scientific cooperation. I decided to acknowledge these collaborations by using the scientific plural throughout the thesis. While Chapter 1, Chapter 2, Chapter 5 and Chapter 6 present original content which was exclusively written for this thesis, parts of Chapter 3 and Chapter 4 are based on co-authored papers which have been published at international peer-reviewed conferences. In addition, some projects were supported by practical works of students which were carried out under my constant supervision and guidance. The following two sections point out the concrete collaborations.

Chapter 3 – Problem Space Exploration

Section 3.2 is partly based on a scientific paper which resulted from a collaboration with the Leibniz Universität Hannover. While Marian Harbach had the original research idea, the elaboration and realization of the research project was collaborative work. The study design and the software was developed by Marian and me and implemented by Andreas Fichtner (student assistant). The user study was simultaneously performed in Munich and Hannover whereby I was responsible for the conduct of the user study at LMU Munich. Finally, Marian and I were equally involved in coordinating and writing the co-authored paper which was published at SOUPS'14 [118].

Section 3.3 is partly based on a scientific paper which resulted from a collaboration with Paul Dunphy from the Newcastle University. The research idea was developed together with Alexander De Luca. While parts of the analyzed raw data originated from a user study performed by De Luca et al. [69], I was responsibility for the planning, implementation and conduction of the presented research project. I implemented the PIN-based prototype and performed the user study. My co-authors gave valuable input for the data analysis and helped with writing the paper which was published at MobileHCI'13 [266].

Section 3.4 is partly based on a scientific paper. While I came up with the original research idea, the study design was developed in collaboration with the co-authors of the later paper. In addition, the research project was supported by Philipp Janssen who implemented the study software as part of his bachelor thesis [139]. I was primarily responsibility for the software architecture and the analysis of the presented data. I was the leading author of the co-authored paper which was published at CHI'15 [265].

Section 3.5 is partly based on a scientific paper. In addition, the analyzed data was collected as part of a bachelor thesis by Peter Arnold [16] and the clustering software was implemented by Iris Maurer and Sascha Oberhuber as part of a seminar on scientific working and teaching in 2014 [172]. I came up with the original research idea and developed the concept of the similarity metric. In addition, I was responsible for the study design and the data analysis. Large parts of the presented data were exclusively analyzed for this thesis. Finally, I was the leading author of a co-authored paper which will be published at MUM'16 [267].

Chapter 4 – Design Space Exploration

Section 4.2 has partly been published as a scientific paper. I came up with the original research idea and was mainly responsible for the concept development. In addition, I was responsible for the design of the performed user studies and analyzed the data presented in this thesis. The practical execution of the first user study was supported by Anton Koslow as part of his bachelor thesis [157]. Anton Koslow implemented the software prototypes and helped with the lab evaluations. The results of the first design iteration were published at IUI'13 [268]. While I was the leading author of the paper, my co-authors gave valuable input and helped structuring the paper. The second development cycle was supported by Alexander Kehr as part of his bachelor thesis [144]. Alexander Kehr implemented the study software and helped with practical execution of the of the user studies.

Section 4.3 is partly based on three scientific research papers. Parts of XSide resulted from collaborations with the Universitella Svizzera italiana and the Leibniz Universität Hannover. However, Alexander De Luca and I were primarily responsible for the concept development, study design and data analysis. The project was supported by Max Maurer, who implemented the software and Huong Nguyen, who helped with the practical execution of the first two user studies. In addition, the Università della Svizzera italiana provided valuable data by replicating the main user study. The second iteration was based on a collaboration with the Leibniz Universität Hannover. I was heavily involved in the redesign of the concept, the design of the user study and the data analysis. The user study was simultaneously conducted in Munich and in Hannover whereby I was primarily responsible for the sessions in Munich. While the first part of the project was published at CHI'13 [74], the second part was published at CHI'14 [71]. I co-authored both papers and contributed large parts of the text.

In addition, parts of the SwiPIN project have already been published as scientific paper. I had the original research idea and was responsible for the concept development, the study designs and the data analysis. Preliminary evaluations were supported by Annika Busch as part of her bachelor thesis [48]. The further development was supported by Bruno Brunkow as part of his bachelor thesis [43]. Bruno Brunkow helped with the practical execution of the second lab study and implemented the required software. The results of the lab evaluations were published at CHI'15 [262]. While I was the leading author, the co-authors gave valuable input on the structure of the paper. Finally, Miriam Mickisch [182] helped with the practical execution of the field study. The task was performed as part of a practical research project.

Section 4.4 was supported by a bachelor thesis, a master thesis and a practical research project. I came up with the original research idea and was responsible for the concept development, the study design and the data analysis. The development process of the background schemes was supported by Malin Eiband as part of her master thesis [88]. Malin Eiband implemented the study software and helped with the data collection. The evaluation of the presentation effects was supported by Anna Kienle, who implemented the study software and helped with the data collection [148, 149]. Finally, parts of the section have been included in a co-authored paper which will be published at MUM'16 [267].

ACKNOWLEDGMENTS

It is time to say “thank you” to my supervisors, my colleagues (aka. friends) and my family who have accompanied my path and who have always supported me throughout the process. I could possibly fill another 200 pages with listing all the received support, and with anecdotes from all the joint experiences we have had. However, this would generate unnecessary printing costs. Therefore, I opted to streamline the process by **using your names to derive keywords**. These keywords, which can be created from the letters of your names, served as a basis for my acknowledgments.

I like to thank my supervisor, **Heinrich** Hußmann, who made this thesis possible in the first place by always supporting and refining my ideas and by striking the right balance between freedom of action and guidance. Your keyword is *human*, because you always ensure that the team is composed of friendly and open-minded human beings and not of workaholic individualists. This strategy enables a cooperative and trustful work environment that is hardly found anywhere else. In addition, I like to thank my external supervisors, Matthew Smith and Andrea Bianchi. **Matthew**, your keyword is *“twist them”*, because that was the first advice you gave me in 2013 after I had presented our work on Back-of-Device Authentication. I am looking forward to our future collaborations. **Andrea**, you were one of the first to *enrich* authentication on mobile devices and you inspired a lot of my work.

Furthermore, a big thank you goes out to my former and current colleagues at the LMU Munich. You are great and each and everyone of you contributed to the awesome time I had during the last five years. **Alexander** De Luca, you have introduced me to the world of usable security. I cannot thank you enough for all your support, which *excelled* all my expectations. It started with the project thesis and continued with years of fruitful collaboration and friendship. Thank you for believing in my skills strongly enough to even pay the first months of my contract at LMU.

What would we all do without our two good souls who keep MiMuc running. **Rainer** Fink, you would never *refrain* from helping us out. Thank you for all the technical support and your unmistakably sense of humor. **Franziska** Schwamb, you are much more than the *bank* which I paid too seldom. Thank you for all the help with diverse bureaucratic obstacles and thank you for your forthright comments which always made all but one laughing.

Florian Alt, I admire how you *ration* your time to collaborate in so many projects at once. **Daniel** Buschek, thank you for all the discussions which *unleashed* new concept ideas and improved my understanding of machine learning. **Andreas** Butz, thank you for your great input at the IDCs and thank you for motivating me to *run* (at least 6.1 KM). **Malin** Eiband, thank you for sharing your road-trip hits which really *beam*ed me back into middle ages. Keep on track and protect our visual privacy! **Maria** Fysaraki, thank you for calling me an extreme athlete because I *ski*. **Ceenu** George, you are a true enrichment for the team and your *urge* for knowledge is exemplary. **Renate** Häuselschmid, we can all be glad that you decided to leave IAV despite your *car enthusiasm* to join us as a full time member. **Alina** Hang, your statements always hit the *nail* on the head and I already miss your philosophical treatises.

Doris Hausen, with you in the office there was no rumor, which remained *unheard*. We were already colleagues twice and I wonder how our paths will cross in the future? **Fabian** Hennecke, the *chief* of wordplay: Thank you for introducing me to the art of Nerf guns modding. **Axel** Hösl, you taught me that the "high five" is the *hello* of modern men. Thank you for all the great conversations from the lowest to the highest level. **Mohamed** Khamis, thank you for giving Erasmus requests a new *home* and thank you for the most perfectly organized IDC ever. **Sebastian** Löhmann, we *established* a nice tradition, let us visit even more live concerts in the future. **Christian** Mai, your keyword is *charisma* because you are the most charismatic food critic I know. **Max** Maurer, thank you for the Nintendo sessions in Cape Town and thanks for enabling the *alumnae* to keep in touch. **Henri** Palleis, thanks for your deadpan humor and your unique coolness. I am impressed how you can *reshape* a dull IDC talk into a humorous performance. **Bastian** Pfleging, your keyword is *enabling gifts* because I am impressed by your persuasive phone calls. If I will ever have trouble with the Deutsche Bahn, I will call you. **Hendrik** Richter, thank you for teaching me how to remain a *"rich drinker"* by becoming the "Milchholer des Monats". **Hannah** Schneider, thank you for joining all the fun in the office and outside of the office and thanks for the *enhanced* watermelon. **Tobias** Seitz, thank you for accompanying my weird *oasis* song interpretations with your guitar and thank your for completing the A206-crew. **Bernhard** Slawik, thank you for presenting the most extraordinary *hardware* prototypes. **Simon** Stusak, thank you for the great *amount* of free drinks at your after-show parties and thank you for hosting our undergraduate learning group. **Sarah** Tausch, thank you for being such a kind and honest colleague, I am already missing our *chats* in the coffee kitchen ... and I am still waiting for your breakdance performance! **Julie** Wagner, thank you for your constant *renewal* of social activities. I really hope, we can establish an annual "Hüttengaudi". **Alexander** Wiethoff, thank you for spreading the "relaxed" vibes. Finally, I like to thank all the great people which could not be addressed directly like the motivated **students** which have supported my projects in over 35 bachelor theses, master theses and practical research projects.

In addition to the great team in Munich, I like to thank all the other colleagues from around the world that I have met during the last five years. Two collaborations shall be specifically highlighted: **Marian** Harbach, due to a *chain* of fortunate circumstances we started fruitful collaborations and what's more, ski trips! **Paul** Dunphy, thank you for supporting me with writing my very first "pattern paper" and thank you for *haply* running with me into the best night ever. I hope to stay in touch with all of you!

The last paragraph is dedicated to my family. Thank you **Sylvia** and **Christoph** von Zezschwitz. You are the best parents I can even imagine. This thesis would not have been possible without all your support and love during my childhood, school days and student life. I cannot express how grateful I am. Thank you **Sophie** von Zezschwitz, the best sister on planet earth! Your believe in my strengths actually made me stronger. Finally, I want to thank **Stefanie** Stich. Stefanie, I love you so much. Thank you for all the love and all the support and sorry for the mood swings and the long working hours.

Let's discover the world!

TABLE OF CONTENTS

LIST OF FIGURES

List of Tables

I

Introduction & Background

Chapter 1

Introduction

We are stuck with technology when
what we really want is just stuff that works.

– Douglas Adams, Writer (2002) –

In 2014, “The Independent”[1] claimed that “there are officially more mobile devices than people in the world”. Indeed, the number of mobile devices has massively grown over the last decade. Especially the smartphone has become our everyday companion. We use it to play games, to communicate with friends or to perform bank transactions. With reference to Douglas Adams’ quote, it finally feels like we have got rid of technology and use “stuff that works”. Unfortunately, the easy usage of mobile devices is only one side of the coin. Smartphones store a lot of personal data and despite all the technological progress, the usability and security of mobile authentication mechanisms did not significantly evolve. As a consequence, data protection is still an annoying and sometimes difficult task which requires an active contribution of the user. This thesis investigates how authentication can be tailored to mobile devices to transform it from “technology” to “stuff that works”.

This Chapter introduces the research topic and provides an overview of the thesis. After a general introduction (Section 1.1), Section 1.2 motivates the efforts taken to improve usability and security of authentication on mobile devices. In addition, we describe the research approach (Section 1.3) and outline the main contributions (Section 1.4). Finally, Section 1.5 presents the structure of the thesis and gives an overview of its content.

[1] `http://www.independent.co.uk/life-style/gadgets-and-tech/news/there-are-officially-more-mobile-devices-than-people-in-the-world-9780518.html` – last accessed: 2016/08/06.

1.1 Mobile Devices and Authentication

Before we investigate the risks and potentials of graphical and gesture-based authentication for touchscreen mobile devices, we introduce general aspects of authentication and mobile devices and set the scope of this thesis. First of all, we define the term "mobile devices" and discuss the strengths and weaknesses of such devices. Secondly, we give a short introduction to authentication and illustrate available authentication factors. Finally, we motivate the investigation of graphical gesture-based secrets and explain why we assume that such concepts have the potential to enable usable and secure authentication on mobile devices.

Strengths and Weaknesses of Mobile Devices

The NIST Special Publication 800-124 [233] defines mobile devices by having a (a) "small form factor", by providing (b) "at least one wireless network interface", by having a (c) "local built-in (non-removable) data storage", by using (d) "an operating system that is not a full-fledged desktop or laptop operating system" and by supporting (e) "applications available through multiple methods". Our definition of mobile devices builds upon the definition of the National Institute of Standards and Technology. However, as we are specifically focusing on mobile devices which support touch-based user interaction and store personal data, we define that mobile devices considered in this thesis provide (f) a visual display with a touchscreen and are primarily used to (g) manage personal data or business data.

The definition excludes laptops, devices with minimal computing power (e.g., cell phones) and devices with a limited scope of application (e.g., handheld game consoles). However, it includes smartphones, tablet computers and touch-based wearable computers (e.g., smartwatches). While the research covered in this thesis is relevant for all three mobile device classes, we opted to focus on the most prominent representative, that is, the smartphone. In 2007, Apple released the first multi-touch enabled smartphone[2]. Eight years later, Google announced that 1.4 billion people are using smart Android devices[3]. In 2016, the Ericsson Mobility Report announced 3.4 billion smartphone subscriptions worldwide[4]. Such numbers show that mobile devices have become ubiquitous and people started living in a mobile society. While a decade ago, cell phones have been mainly used for calling and texting, phones of 2016 are universal computing devices with almost endless application areas [201]. They are used to take photos, to carry out banking transactions and to surf the web. Mobile Internet enables users to edit, store and access a myriad of data anywhere and anytime. The development of the Internet of Things [110] is likely to drive this process forward and create an interconnected world of even more mobile devices and more opportunities.

[2] `http://www.apple.com/de/pr/library/2007/01/09Apple-Reinvents-the-Phone-with-iPhone.html` – last accessed: 2016/08/06.

[3] `http://www.ubergizmo.com/2015/09/over-1-4-billion-people-are-now-using-android/?utm_source~=~mainrss` – last access: 2016/08/02

[4] `https://www.ericsson.com/mobility-report/mobile-subscriptions` – last access: 2016/08/02

However, the use of smart mobile devices introduces several security and privacy threats. The user's digital life is often either directly stored on the device (e.g., photos) or accessible through cloud-based services (e.g., e-mails). In addition, mobile devices are carried around and face greater risks of getting lost or stolen than stationary home computers. Finally, mobile devices are often used in public spaces where user interaction is easy to observe and unauthorized access is more likely. Authentication often represents the only countermeasure to protect unattended devices from unwanted access. However, the specific form factor of mobile devices and the mobile context pose special demands on feasible authentication mechanisms. It has already been shown that many users underestimate the risks and do not accept putting extra effort in using secure lock screens [84]. In addition, even if authentication methods are used, established concepts are easy to attack [254, 262], not usable on mobile devices [178] or demand the exposure of personal information [117]. Section 1.2 discusses these problems in more detail. We argue that adequate usability can only be achieved if authentication methods are tailored to the specific needs of mobile device users. Instead of transferring existing methods from other authentication scenarios (e.g., ATMs), developers need to consider real-world user behavior and understand real-world threats. The next Section gives a brief introduction to the authentication factors which can be exploited to build authentication mechanisms for mobile devices.

A Brief Introduction to Authentication

Authentication, the act of confirming the identity of a person or a document, has been important long before computers existed. For example, Ali Baba and the Forty Thieves[5], which was published over 200 years ago, tells the story of Ali Baba, who gains access to a cave by using the words "open sesame". Technically speaking, the tale describes the use of a password-based authentication system. Over 50 years ago, authentication was introduced to computers [62]. The MIT Compatible Time-Sharing System (CTSS) was the first computer system which requested a text-based password [63]. Over the next decades, passwords were only required for specific application areas and used by computer experts. However, with the broad introduction of home computers in the 1980s and the World Wide Web in the 1990s [190], password-based authentication became part of most people's everyday life. Finally, with the introduction of smartphones, authentication has entered the mobile context.

Authentication methods can be categorized according to the used authentication factor. In general, we distinguish knowledge-based, token-based and biometric concepts [196]. Knowledge-based authentication can be referred to as "something the user knows" [196]. While text-based passwords are the most prominent representatives, several alternative concepts (e.g., graphical passwords) have been discussed over the last two decades [32]. In the context of mobile devices, PINs and gestures [266, 295] have become widely accepted. Token-based concepts authenticate users based on the possession of certain objects (e.g., a physical key). Such concepts are also referred to as "something the user has" [196]. Representatives of this class are widely deployed in form of bank cards which need to be provided

[5] `http://www.bartleby.com/16/905.html` – last accessed: 2016/08/03.

to authenticate with automatic teller machines. In the mobile context, the “NFC ring” was proposed as a commercially available product[6]. However, even though the vendor claims that “it couldn’t be easier”, token-based mechanisms are still uncommon in the context of mobile devices. Finally, biometric authentication is based on the analysis of physical features or behavioral characteristics. It is often referred to as “something the user is” [196]. Based on the considered cues, biometric systems are distinguished into physical or behavioral approaches. While the latter is not suited for ad-hoc authentication scenarios (i.e., mobile lock screens), it was already proposed as a second layer of security [69]. In the mobile context, biometric approaches have been introduced in form of face recognition and fingerprint recognition [70].

The next Section outlines the advantages and disadvantages of the presented authentication factors and illustrates that knowledge is still an indispensable authentication factor.

Knowledge: An Indispensable Authentication Factor

The previous Section presented the basic authentication approaches. While each authentication factor has specific advantages and disadvantages, we claim that knowledge-based approaches are still indispensable.

Tokens are easy to use and require little effort. However, they are prone to be forgotten or stolen [196]. This is especially true in the mobile context, which requires security tokens to be carried around. As the possession of a security token is sufficient to authenticate its owner, such concepts usually require a second security factor. For example, authentication on automatic teller machines requires a bank card (token) and a PIN (knowledge) [134]. We therefore argue that tokens are not suited as standalone authentication factor. However, they can complement other factors to improve the overall security and usability of an authentication system [36].

Biometric authentication mechanisms like fingerprint recognition are easy to use and require little effort. This is especially true, when fingerprint readers are integrated into buttons which would be activated anyway[7]. However, there are fundamental arguments against the extensive use of biometric features. Firstly, physical biometric information represents highly personal data which is directly liked to the user. On the one hand, this means that the biometric information could be used to identify users in any other context. On the other hand, this means that users cannot change the information if it falls into the wrong hands. Secondly, it has already been shown that biometric cues are easy to fake [8]. As Frank Rieger, a security expert and hacker of the Chaos Computer Club (CCC) puts it, “It is plain stupid to use something that you can’t change and that you leave everywhere every day as a security token” [9]. However, even if we ignore these threats and assume that the biometric data will be

[6] `http://store.nfcring.com` – last accessed: 2016/08/03.

[7] `https://support.apple.com/en-us/HT201371` – last accessed: 2016/08/21.

[8] `http://www.ccc.de/en/updates/2014/ursel` – last accessed: 2016/08/21.

[9] `https://www.ccc.de/en/updates/2013/ccc-breaks-apple-touchid` – last accessed: 2016/08/21.

securely stored on the device, biometric concepts are hardly suitable as standalone solution. As the recognition of biometric features is prone to errors, biometric authentication requires a second factor as fallback solution [114]. For example, Apple's TouchID[10] is combined with a knowledge-based authentication system and requires the user to enter a passcode after (a) restarting the device, if (b) fingerprint recognition failed five times or if the (c) device was not unlocked for more than 48 hours. We argue that biometric solutions can complement knowledge-based authentication mechanisms but they are unlikely to replace such concepts. For example, behavioral cues can be used as a second factor to improve the overall security and usability of authentication on mobile devices [137].

Even though knowledge-based authentication mechanisms are far from perfect as they demand physical and mental effort and are prone to several attacks [18, 38, 207], they remain indispensable [247]. Nevertheless, we argue that knowledge-based concepts need to be tailored to the respective application area to enable usable and secure authentication. In the context of smart mobile devices, knowledge-based authentication must not be restricted to pressing buttons. In contrast, authentication concepts need to exploit the specific features of mobile devices which provide rich input and output capabilities. In the scope of this thesis, we specifically focus on graphical gesture-based solutions and investigate how mobile authentication can be improved in terms of usability and security.

1.2 Problem Statement and Research Questions

As already indicated in Section 1.1, unwanted access on mobile devices can have serious consequences for the device owner. Since smart mobile devices can store any kind of personal information or business-related data, the threat model comprises a wide range of risks. To classify such risks, we can distinguish between different attacker types [279], different attack scenarios [269] and different consequences [118]. Since applications are usually not protected, device access enables attackers to steal data, to modify data or to perform actions which cause damage in the physical world (e.g., identity theft). In order to reduce the risk of unwanted access, mobile devices can be protected by secure lock screens. As outlined in Section 1.1, such lock screens require users to authenticate before access is granted. Besides PIN and biometric solutions, gesture-based unlock mechanisms have gained acceptance. The most prominent representative of this type of authentication is the Android pattern unlock [18]. In 2015, the mechanism was available on 83% of all smartphones[11].

In theory, such unlock methods provide high security at low cost. In the case of unlock patterns, the user would select one out of 389.112 [18] available gestures, ideally one that is quick and easy to perform. Since mobile devices limit the number of failed authentications[12], a potential attacker would have very little chance to guess the right gesture to unlock

[10] https://support.apple.com/en-us/HT201371 – last accessed: 2016/08/06.

[11] http://www.idc.com/prodserv/smartphone-os-market-share.jsp – last accessed: 2016/08/04.

[12] https://support.apple.com/en-us/HT204306 – last access: 2016/08/06.

the device. However, the reality is different. Firstly, many users claim that secure lock screens are too cumbersome and deliberately sacrifice security for usability. Secondly, even if lock screens are used, security is not guaranteed as the theoretical protection of current solutions is jeopardized by various real-world factors. For instance, password selection is predictable [38] which makes user-chosen secrets easy to guess. In addition, the input of current secrets is easy to observe and makes so-called shoulder surfing attacks effective. This is particularly critical in the mobile context where shoulder surfing is facilitated by uncontrolled environments. Finally, touchscreens enable so-called smudge attacks [18], a threat which is based on the analysis of oily residues which are left on the display.

As these threats are known for several years [32], alternative concepts have already been proposed to increase the practical security. Most of these mechanisms aim at protecting user input from observation attacks (e.g., [27, 246, 258, 278]). Other methods try to increase guessing-resistance (e.g., [83, 232, 251]). However, despite the numerous efforts made in terms of improving security, none of the methods achieved a breakthrough. Section 2 gives a detailed overview of previous work. The literature review indicates that usability is often the main limiting factor. Since the mobile context makes great demands on usability and the use of secure authentication mechanisms is optional, user acceptance can only be achieved by designing satisfying authentication mechanisms which are tailored to the mobile context. However, many real-world aspects of mobile authentication mechanisms are still unknown. This is especially true, when considering graphical gesture-based authentication. In addition, the development of novel authentication mechanisms is often characterized by explorative approaches and lacks comparable design and evaluation processes.

Derived from the problems stated above, this thesis aims at providing in-depth knowledge on the risks and potentials of graphical gesture-based authentication on mobile devices by answering the following main research questions:

RQ1 How do established gesture-based concepts perform in terms of usability and security [current state]?

RQ2 What are the requirements for improved authentication on mobile devices [goal state]?

RQ3 How must graphical gesture-based concepts be designed (and evaluated) to meet the requirements of mobile devices [process] ?

1.3 Research Approach

The research problem is investigated based on an empirical approach involving both basic research and applied research. While the basic research aims at gathering an in-depth understanding of the authentication context and the impact of specific design and evaluation strategies, the applied research focuses on the development of feasible authentication mechanisms for mobile devices.

Following the recommendations by Goel and Pirolli [108], we structure authentication on mobile devices as a design problem space. Consequently, we can address the research problem by investigating the interconnection of the design space and the problem space. The problem space of mobile devices comprises usability (e.g., efficiency) and security (e.g., observation resistance) aspects. While the current state of the problem space is given by the status quo of mobile device authentication, its goal state will be defined in this thesis and represents the ultimate research goal. The design space is given by the set of design factors which describe a graphical gesture-based authentication system.

Based on the framework of triangulation by Mackay and Fayard [168], the research questions are addressed in different ways. Firstly, the problem space is explored to define the current state of authentication on mobile devices. The basic research involves explorative and descriptive studies and gathers qualitative and quantitative data. We perform natural and controlled field studies to analyze the usability of currently used methods and to understand the context of mobile authentication. In addition, we perform online experiments to investigate the security of currently used gestures and describe their guessability and observability.

The results of the problem space exploration are complemented by a systematic analysis of the design space. The design space of graphical gesture-based authentication is investigated following an experimental design-oriented research approach [91, 297]. Based on a user-centered design process [193], we develop various concepts which are intended to solve specific subproblems in isolation [108]. The iterative development process involves focus group discussions, low-fidelity prototypes and interactive high-fidelity prototypes. The concepts are evaluated in controlled lab studies and field studies to prove their usability and security. The explanatory research gathers qualitative and quantitative data and provides valuable insights into the impact of specific design decisions.

Finally, the insights of the problem space exploration and the design space exploration are combined to describe the interconnection of the design factors and the usability and security of graphical gesture-based authentication mechanisms.

1.4 Main Contributions

The presented work contributes to the field of usable security by providing both fundamental insights and practical solutions. Firstly, we contribute to the understanding of the current state of mobile authentication. The thesis presents important insights into authentication context, user behavior and the usability and security of currently deployed systems. Secondly, we present practical solutions in form of novel authentication mechanisms which were proven to be more secure but yet usable in the mobile context. Finally, we provide generalized insights into the impact of specific design factors and present recommendations for the design and the evaluation of graphical gesture-based authentication mechanisms. In the following, we give an overview of the main contributions.

The Current State of (Gesture-based) Mobile Authentication

In addition to an extensive literature review which provides insights into the current state of research, we investigate the real-world demands of mobile authentications. For this purpose, we specify the problem space of gesture-based authentication on mobile devices and discuss its goal state. In addition, we provide valuable insights into the usability and security of currently used authentication methods. We analyze the real-world behavior and risk perception of smartphone users and investigate the usability and security of current gesture-based solutions in the field. In addition to the qualitative and the quantitative findings, we present theoretical models which describe the usability and the security of currently used concepts. In this regard, we present a taxonomy of gesture-based input errors, a prediction model for the observability of grid-based gesture input and a novel metric to assess the practical password space of grid-based gestures. Overall, the in-depth description of both real-world factors and essential requirements will assist in developing new solutions which are particularly tailored to the context of mobile devices.

Feasible Authentication Methods for Mobile Devices

We present a systematic exploration of the design space of graphical gesture-based authentication. For this purpose, we firstly specify important design factors and discuss their main characteristics. The investigation is based on a design-oriented research approach which results in several practical solutions for existing real-world problems. We analyze how gesture-based authentication can be improved to prevent smudge attacks, observation attacks and guessing attacks. The novel authentication concepts are implemented as working prototypes and evaluated in the lab and in the field. The results indicate that the presented concepts are significantly more secure than current solutions and provide good usability in the context of mobile devices. In addition to the presentation of concrete solutions, we contribute by discussing general implications of the effects of both specific design decisions and evaluation strategies.

The Interconnection of Design and Problem Space

The results of the basic research and the design-oriented applied research are combined to discuss the interconnection of design space and problem space. First of all, we contribute by outlining crucial requirements for mobile authentication mechanisms and present ten empirically proven objectives for future designs. The design objectives will help researchers and designers in defining concrete goals for their projects. In addition, we provide an in-depth discussion of the observed interaction effects between design factors and problem space. The insights are presented in a way that enables a systematic goal-directed development process for mobile authentication mechanisms. Finally, we outline two potential application areas for such a process and show how the insights of this thesis can facilitate both bottom-up design approaches and top-down analysis. Overall, the presented insights contribute to the field by providing assistance to researchers and developers who design and evaluate mobile authentication systems. Moreover, the applied methodology is in principle transferable to very different research questions, in particular in the area of usable security and privacy.

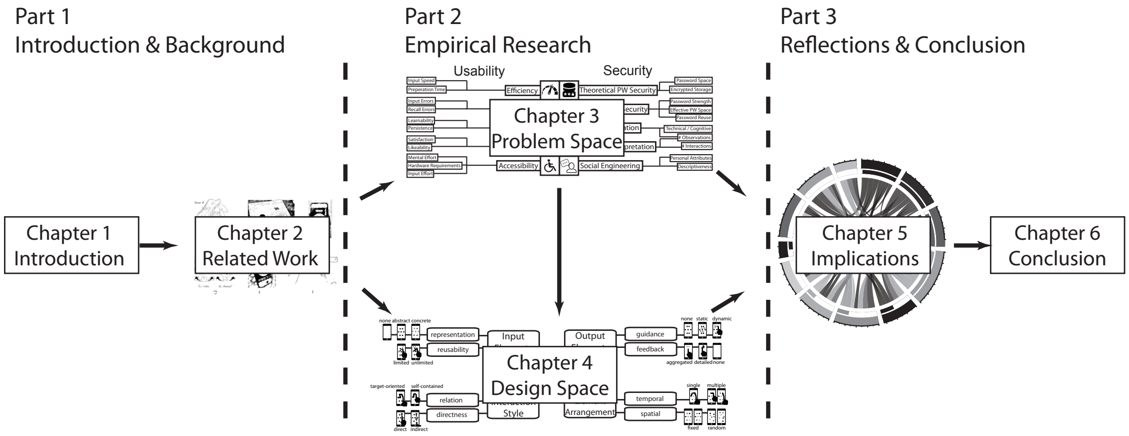


Figure 1.1: The thesis comprises six main chapters. After an introduction to the research problem, we investigate the problem space and the design space. Chapter 5 discusses the implications of the results before Chapter 6 concludes the thesis.

1.5 Thesis Structure

Figure 1.1 illustrates the structure of the thesis which is organized in three parts and six chapters. *Chapter 1* introduces the research problem and the research approach. *Chapter 2* provides an extensive review of previous work and identifies open questions. *Chapter 3* and *Chapter 4* provide an in-depth investigation of authentication on mobile devices to bridge the identified gaps in knowledge. While *Chapter 3* presents field studies which analyze the current state of authentication on mobile devices, *Chapter 4* explores how the current state can be improved. *Chapter 5* combines the insights of *Chapter 3* and *Chapter 4* and provides general recommendations for the design and evaluation of feasible authentication mechanisms. Finally, *Chapter 6* concludes the thesis. In the following, we summarize the content of the remaining chapters in more detail:

Chapter 2: Related Work This Chapter presents an extensive review of related work. Section 2.1 presents the most important insights of traditional usable security research. It discusses what we can learn from text-based passwords and illustrates crucial findings concerning user behavior, password management, strength metrics and password selection. Section 2.3 illustrates the evolution of graphical and gesture-based authentication mechanisms which were proposed as an alternative to text-based passwords. It introduces common approaches and discusses their general strengths and weaknesses. Consequently, Section 2.3 presents several research projects which aim at improving the security and usability. The presented concepts illustrate how graphical elements and gesture-based interaction can be used to prevent guessing attacks, smudge attacks and observation attacks. Finally, Section 2.4 summarizes the lessons learned and points out open questions.

Chapter 3: Problem Space This Chapter presents four research projects which explore the problem space of gesture-based authentication. Section 3.1 provides a definition of the problem space and gives an overview of the research covered. Section 3.2 presents a natural field study which investigates unlocking behavior and risk perception in the wild. Section 3.3 investigates the usability of grid-based gestures by comparing PIN and unlock gestures in the field. It presents important insights into the real-world usability of established concepts and provides a taxonomy for gesture-based input errors. Section 3.4 analyses the observation resistance of an established gesture-based authentication mechanism. It presents the results of an online experiment which indicate that established grid-based authentication concepts are prone to observation attacks. In addition, it provides a prediction model which can be used to assess the observability of a given gesture. Section 3.5 investigates the practical password space of grid-based gestures. It presents a novel metric to quantify the similarity of a given gesture set and indicates that user-chosen secrets are very predictable. Finally, Section 3.6 aggregates the acquired knowledge and discusses the implications for future developments.

Chapter 4: Design Space This Chapter presents a systematic exploration of the design space of gesture-based authentication. Section 4.1 defines the design space and illustrates the individual design factors. In addition, it gives an overview of the research projects which are grouped according to the considered threat model. Section 4.2 investigates the prevention of smudge attacks and presents spatially randomized authentication mechanisms which hamper the interpretation of smudge traces. Section 4.3 addresses the problem of observation attacks and illustrates how gestures can be utilized to design usable authentication mechanisms which allow users to increase security on demand. Section 4.4 investigates if the practical password space of gesture-based authentication can be increased through implicit nudging effects. The results indicate that visual guidance can effectively influence gesture selection. Finally, Section 4.5 summarizes the results.

Chapter 5: Implications This Chapter discusses the implications of the interconnection of design space and problem space. Section 5.1 revisits the problem space and discusses the requirements of usable and secure authentication on mobile devices. In addition, it suggests ten concrete design objectives for future developments. Section 5.2 provides design assistance for gesture-based authentication mechanisms on mobile devices. We map design factors and design objectives and discuss observed interaction effects between design space and problem space. In addition, we provide recommendations for or a goal-oriented design process. Finally, Section 5.3 addresses the evaluation of mobile authentication and presents the lessons learned according to exploratory, descriptive and explanatory approaches.

Chapter 6: Conclusion This Chapter summarizes the thesis and discusses future work. Section 6.1 provides a short summary of the thesis and points out the main contributions. Section 6.2 illustrates directions for future work and discusses how the presented insights can be useful in other contexts. Finally, Section 6.3 concludes the thesis with a visionary outlook on authentication on mobile devices.

Chapter 2

Background and Related Work

Study the past if you would define the future.

– Confucius, teacher and philosopher (551 BC - 479 BC) –

In 1975, Saltzer and Schroeder argue that “It is essential that the human interface be designed for ease of use, so that users routinely and automatically apply the protection mechanisms correctly [..]" [209]. This claim can be seen as the starting point of a whole research area concerning usable security and privacy [106]. Over the last decades, usable security research kept constantly growing and a nearly unmanageable amount of research papers was published. This Chapter presents the most relevant work, identifies unsolved problems and discusses how these gaps can be filled.

Section 2.1 presents results concerning text-based passwords: We will learn that usability features have a massive impact on the security of the respective authentication system. The discussion sheds light on user behavior in the wild and password management strategies. Motivated by the usability issues of text-based passwords, research began designing alternative authentication mechanisms: Graphical and gesture-based authentication systems. Section 2.2 illustrates the evolution of such systems and presents the general approach. The discussion will show that usability and security problems are not automatically solved by eliminating text-based secrets. On the contrary, authentication times are often high and practical security is jeopardized due to increased observability. As a consequence, Section 2.3 focuses on improving the practical security of gesture-based graphical passwords. We will learn that design decisions have a significant impact on the performance of an authentication system. Finally, Section 2.4 summarizes the insights gained from this Chapter and discusses both lessons learned and open questions.

2.1 Learning from Text-based Passwords

The basis for the thorough design of improved authentication concepts must be a profound understanding of existing problems and solutions. This Section gives important insight into user behavior and password selection strategies in the context of desktop computers. As most of the fundamental research presented in this Section is concerning text-based passwords entered with a physical QWERTY keyboard, not all results can be directly transferred to the context of mobile devices. However, the knowledge gained from this Section will both motivate the effort and help with the development of usable and secure concepts which are tailored to touch-based mobile devices.

2.1.1 User Behavior and Password Management

User behavior regarding alphanumeric passwords has been extensively investigated for over 20 years [2]. It was shown early that users are overstrained by the large number of required passwords [1, 212]. In 2014, an average user had to manage 27 password protected online accounts [238]. It is obvious that memorizing a unique and complex alphanumeric string for each of those accounts would significantly exceed the capacity of the human brain [1, 98, 194]. Over the years, the password problem becomes worse as the number of required secrets increases with every registered service [194].

While a few users simply think that hacking is not a problem [2], Herley and van Oorschot [129] note that most users are indeed aware of security risks. Still, the direct usability costs of using unique and complex passwords are usually much higher than the indirect threat of potential attacks [126]. As a consequence, people start optimizing usability by reusing passwords [107, 133, 205], by writing down passwords [132, 135, 212, 286] and by sharing their secrets with other persons [1, 143, 225].

Password Reuse

Password reuse is the main coping strategy for most users [95, 109, 225]. Gaw and Felten [107] found out that the number of reused passwords positively correlates with the number of protected accounts. Stobert and Biddle [238] report that 96% of the participants reused passwords. More precisely, they used a median of five passwords to manage 27 accounts. 88% of the participants stated to reuse more than one password and 73% acknowledged to reuse passwords frequently. However, users do not randomly reuse passwords but cluster their accounts into different categories [98, 205, 238]. For this purpose, users usually hold a (primitive) primary password for most of their accounts and use more complex secrets for sensitive services [205,238,263]. Unfortunately, the same composition strategies are used to select low security and high security passwords [225] and therefore these primary passwords are good predictors for more complex ones [263]. In addition, von Zezschwitz et al. [263] revealed that such composition strategies remain stable over time as new passwords often have roots in the very first passwords ever chosen.

Ives et al. [135] published one of the first works which pointed out the security risk of password reuse and argued that no account can be more secure than the weakest one using the same secret. Das et al. [65] quantified this risk and presented a cross-site password-guessing algorithm which exploits password reuse for attacks. The algorithm was able to guess 30% of the transformed passwords within 100 attempts. This was significantly better than a compared traditional guessing algorithm. A similar attack was proposed by Haque et al. [116]. Florencio et al. [98] point out that password reuse cannot be avoided even when memorability-optimized passwords are assumed. As a consequence, several researchers [98, 133, 191] postulate the need for usable guidelines which illustrate how to reuse and recycle passwords in an optimal and secure way. While it generally can be assumed that the same reuse behavior exists for (remote) services accessed on *mobile devices*, Egelman et al. [85] recently reported that password reuse is also common for primary unlock screens.

Password Storage Behavior

In 1999, Adams and Sasse [1] published one of the first extensive surveys concerning authentication behavior and reported that all users regularly wrote down their passwords. In the same year, Zviran at al. [298] interviewed 860 password users and found out that 35% of the users wrote down passwords. They criticized that most people store their passwords in unsecure places like wallets, notebooks and calendars. While in 2006, physical record was still reported to be most common [44, 107], in 2014, most people stored passwords digitally [238]. Despite the use of password managers and cookies, the storage behavior is quite diverse. Stobert and Biddle [238] report the use of spreadsheets, cloud services or e-mail accounts. However, physical storage is still common [225, 238].

Komanduri et al. [155] argue that strict password policies are the primary reason for writing down passwords. Zezschwitz et al. [263] found that password storage behavior correlates with the number of used passwords. Zviran and Haga [298] state that password complexity and frequency of use influence storage rates, but length has no significant impact. Komanduri et al. [155] quantified the relationship of password complexity and password storage and found out that users wrote down 17% of the simple passwords, but 50% of the complex ones. Grawemeyer and Johnson [109] performed a seven days diary study and found that the odds of writing a password down were 18 times higher if it was unique, and ten times higher if it was rarely used. As a consequence, research came to the conclusion that writing down passwords can also have positive effects [98, 212]. As long as the passwords are stored in a secure place, it allows users to use a more diverse set of more complex passwords [98]. *Mobile device* use further increases storage rates [86]. As typing on such devices is significantly more cumbersome and error prone than on desktop computers [102, 174, 264], most mobile applications store passwords digitally and remain logged in per default [86, 171]. Therefore, the unlock mechanism of mobile devices often represents the only barrier of protection from unauthorized access to the user's digital life [142].

Password Sharing

While password reuse and password storage represent personal coping strategies which are first of all triggered by the number of required passwords, password sharing is usually influenced by social and organizational factors [116, 143, 274]. Kaye [143] analyzed the sharing behavior of 122 internet users and found out that e-mail and social media passwords were the most shared types when the private context was considered. In the working environment, colleagues often shared passwords for joint workspaces or for shared accounts of commercial sites. In addition to such traditional online accounts, people share passwords in various other cases. PIN codes for bank accounts are commonly shared between partners [143, 230], flat mates share passwords for WiFi access or for Netflix accounts [116] and colleagues share passwords of zipped files [133]. In general, sharing can be distinguished into long-term and ad-hoc instances [133]. Inglesant and Sasse [133] state that especially ad-hoc sharing induces insecure password behavior. For example, users quickly choose secrets based on common dictionary words to protect shared files and then email those passwords to the recipient. Beside organizational factors like shared accounts [109] or higher orders [274], the social context has a significant impact on sharing behavior. Sharing a password is often a sign of trust and someone not sharing a password can be seen as someone who has something to hide [212, 274]. Counteracting such behavior is difficult. Adams and Sasse [1] claim that secure behavior only becomes possible when authentication mechanisms fit into the social context of use.

While password sharing is primarily influenced by social factors, it can cause similar security problems as password reuse. Haque et al. [116] revealed that most users utilize the same composition strategies for shared passwords as for the rest of their passwords. Therefore, shared passwords can expose information on the composition of more secure individual secrets. In the *mobile context*, device sharing has become very common [115, 124, 142]. Mobile devices are shared with friends and colleagues for various reasons. Karlson et al. [142] state that a smartphone is used by up to eleven different guests. Since current mobile devices rely on all-or-nothing access [234], the device owner faces the risk of unintentionally exposing sensitive data each time she shares her device. In addition, it was shown that users intentionally reveal their unlock codes to make device sharing easier [85, 257].

2.1.2 Strength Metrics and Password Selection

Alongside to insecure password management, predictable password selection represents one of the main problems of alphanumeric authentication systems [39]. It was found out early that user-selected secrets are often optimized for good memorability [2, 184, 298]. People choose easy-to-remember secrets which are often based on meaningful dictionary words, birthdays or names. Unfortunately, such memorable passwords are also easy to guess as the selection behavior is predictable and thus, the search space is limited to a fraction of its possible size [37]. This induces a significant difference between the theoretical security of alphanumeric authentication and the practical security in the real world.

The first part of this Section presents various password strength metrics and provides insights into the question what makes a password secure. In the second part, we take a closer look on password composition strategies and discuss externalities which influence password selection. Recent work has shown that *mobile devices* have become one of those externalities and that password input on mobile devices is likely to downgrade the usability and the security of traditional authentication mechanisms [264].

Password Strength Metrics

Password strength is traditionally defined by the chance that a secret is guessed by an attacker. If users chose passwords randomly from the set of given characters, the task of quantifying password strength would be straightforward [97]. When assuming a character set C and a password length of L, the chance to guess a random password would be C^{-L}. However, users do not choose characters randomly, but the set of used secrets is skewed towards a subset of all possible combinations [2]. This makes assessing the strength of a given password an exceptionally hard task [87, 97].

One common approach to estimate the security of a given password is based on information entropy which was formalized by Claude Shannon [223] in 1948. Shannon entropy describes the expected value (given in bits) of information in a received message by measuring the information that is unknown due to random variables [97]. The National Institute of Standards and Technology (NIST) [47] proposed to combine the concept of information theory with the characteristics of password selection to assess the strength of human-chosen secrets against online guessing attacks [97]. According to the NIST guideline, "password strength is determined by a password's length and its complexity, which is determined by the unpredictability of its characters" [213]. The Electronic Authentication Guideline [47] gives concrete estimates for the entropy of user-selected passwords. For example, "the entropy of the first character is taken to be 4 bits [...], the entropy of the next 7 characters are 2 bits per character" and "a bonus of 6 bits of entropy is assigned" when the use of upper-case letters and non-alphabetic characters is forced or the use of dictionary words is blocked.

Entropy-based strength metrics have become the standard for most password policies [96] and have been the basis for various scientific evaluations (e.g., [95, 219, 263]). However, since password entropy is a statistical measure, it does not reflect actual user choice. NIST measurements were already proven to insufficiently estimate the difficulty of guessing user-selected secrets [97, 145, 177]. Weir et al. [272] argue that guessing entropy tends to overestimate the security of most passwords, while the strength of some individual but short passwords is drastically underestimated. For this reason, various researchers claim that entropy is not a valid metric to estimate password strength and propose using guessability instead (e.g., [38, 76, 145, 167, 177, 188, 272]). In 1979, Morris and Thomson [184] published the first analysis based on an educated guessing attack. They attacked 3000 passwords with a basic dictionary containing 250,000 words and were able to expose one third of the password corpus. In 2005, Narayanan and Shmatikov [188] exposed 67.6% of the passwords exploiting the contextual frequency of characters in natural language based on Markov chains. Such Markov modeling techniques can replicate often chosen passwords that are not based on dictionary words and are especially effective in breaking strong passwords [76]. Weir et al. [273] propose to analyze the structure from training data and then apply mangling rules. Such attacks use probabilistic context-free grammars based on known composition strategies and benefit from password reuse and leaked password lists [177]. Joseph Bonneau [38] introduces partial guessing metrics which define password strength based on the number of cracked passwords before one specific password is exposed. Kelley et al. [145] propose to count the number of guesses required to guess a specific password.

In summary, password strength is still very difficult to define [87] as information entropy has very limited value in predicting password strength and the estimated guessability of a password depends on the system setup and the training data [145, 272]. Florencio et al. [97] generally challenge the importance of complex passwords in real life and note that guessing resistance is often irrelevant. They argue that online attacks are usually prevented through automatic lockouts and offline guessing attacks are often unnecessary or impossible. Furthermore, even the strongest password cannot protect from threats like observation attacks, key-logging or social engineering. Still, consent can be seen in the fact that a secure password must not be popular [167, 215], that length is usually more important than complexity [224] and that single meaningful strings should be avoided [131]. In addition, specific recommendations for usable and secure alphanumeric passwords were published. Examples are mnemonic passphrases [159, 286], cognitive or associative passwords [46, 125].

Password Selection

Password selection has been extensively studied for over 35 years [184, 247]. The projects can be distinguished into three different categories [106]: (a) surveys which are based on self-reported data, (b) controlled lab- and field studies where users are asked to select a password (c) and the analysis of real-world data from leaked databases or running systems.

Early studies were mostly based on surveys and focused on the semantic content of passwords. Adams et al. [2] analyzed 139 questionnaires and found out that most users selected names and common dictionary words. In 1999, Zviran and Haga [298] analyzed the answers of 860 computer users of the Department of Defense. They revealed that most passwords were based on six characters and represented meaningful strings. Furthermore, 80% of the participants preferred alphabetic characters while only 0.7% stated to exploit the whole ASCII set. In the following years, several other studies (e.g., [42, 44, 205, 238]) confirmed user preferences for short, meaningful and therefore memorable passwords.

To increase the strength of user-chosen passwords, proactive password checkers (password meters) [33, 255, 287] and composition policies [96, 180, 240] were introduced. The effects of such countermeasures were mostly evaluated in controlled lab- and field studies (e.g., [133, 240, 255, 286]). Proctor et al. [203] claim that a minimum length requirement is more important for password meters than other restrictions. Ur et al. [255] ran an online experiment with 2,931 participants and 14 proactive password checkers. They found out that the pure presence of password meters resulted in significantly longer passwords. In addition, some password meters were able to nudge users to use more numbers, upper-case letters or symbols. However, Egelman et al. [87] argue that such positive effects can only be expected when users are actually forced to change their passwords and when the protected account is perceived to be important. Furthermore, it was shown that most password policies do not increase security [225, 272] but are very likely to create additional burdens on the users [155]. Some policies even reduce the possible password strength by restricting the number of characters or by banning certain symbols [133]. In addition, it was shown that the insertion of special characters, numbers and upper-case letters is predictable [97, 155, 263, 272]. Policy-complying passwords usually start with upper-case letters and end with predictable number sequences (e.g., "123") [224]. Most users employ only a subset of possible special characters and use them as a replacement for similar looking letters (e.g., "@" as "a") [136]. Consequently, most of today's passwords are based on similar variations of meaningful words (e.g., P@ssword123) and remain easy to guess [109, 266].

While controlled user studies give valuable insights into the impact of different policy conditions, they often lack ecological validity [87, 247]. This drawback can be overcome by analyzing real-world passwords of running systems [95] or of exposed databases [38, 170, 219]. Florencio and Herley [95] analyzed the password composition behavior of 544,960 web users and confirmed that most passwords were based on lower-case letters as upper-case letters or special characters were hardly used. Bruce Schneier [219] analyzed 34,000 MySpace passwords and stated that 81% of the passwords were alphanumeric, most of them were based on lower-case letters and a single digit. Malone and Maher [170] analyzed various password lists and found out that simple secrets as "123456" and "password" were very common. They furthermore state that passwords often show features which are specific to the protected website. Joseph Bonneau [38] analyzed a corpus of 70,000,000 passwords and revealed that the entropy in the password space is low and that password composition is hardly influenced by the sensitivity of the protected data. Mazurek et al. [177] showed that experienced computer users choose significantly stronger passwords. In addition, they confirmed that the use of digits, symbols, and upper-case letters increases security, but the insertion of such character classes is easy to predict.

Further literature reviews indicate that even if user-selected passwords are still easy to guess, they became stronger over the last decades [205, 219, 266]. However, recent work revealed that the increasing use of passwords on *mobile devices* is likely to stop this positive trend [264, 290]. Maydebura et al. [174] state that the input of alphanumeric passwords on mobile devices is significantly more error-prone. Nevertheless, password protected services are increasingly used on smartphones [171]. In a recent survey by von Zezschwitz et al. [264], 69% of the participants stated that they already had created alphanumeric passwords on a mobile device. 20% of them acknowledged to have used simpler versions of already known PC passwords. Yang et al. [290] confirmed that mobile devices have a negative impact on the password structure. Users tend to select easy-to-enter passwords which comprise more lower-case letters and significantly less upper-case letters or symbols [264].

2.2 The Evolution of Graphical and Gesture-based Authentication

Section 2.1 illustrated several critical issues of alphanumeric passwords. Most of these problems result from the fact that humans are not able to memorize a great number of cryptic alphanumeric strings. For this reason, researchers started looking for alternative secrets that better support the specific skills of the human brain. While most humans might struggle with remembering random strings, the ability to recognize previously seen images and to reproduce trained gestures is widely spread.

In 1996, Blonder [35] was the first to come up with the idea of utilizing pictures instead of alphanumeric characters. In the following years, various graphical and gesture-based password schemes were proposed. Until today, all published methods rely on at least one of the following principles [32]: (a) recognizing objects of a challenge set, (b) localizing specific areas within a given picture or (c) recalling a shape or a gesture. The first two classes are primarily relying on the findings that humans perform better in remembering pictures than words (e.g., [189, 226, 235, 236]). In contrast, concepts of the third category are mainly exploiting the effects of motor learning [3, 94, 221] which indicate that humans are exceptionally good in memorizing consistently repeated movements. As a consequence, some concepts of the latter category actually completely forgo visual feedback [243]. In addition to motor memory effects and the pictorial superiority effect, all three categories can support dual-coding [175, 198, 237]. According to the dual-coding theory, information recall is improved when visual and verbal aspects were used for memorization. This might for example be the case, when a picture illustrates an object which can be named (e.g., a car).

Several user studies have indicated that graphical passwords are easier to remember [32, 104, 183] and that users are in favor of these concepts [53, 68]. According to security, graphical passwords were shown to be less vulnerable to brute-force attacks [32, 104] but password selection remains predictable [66]. In addition, the input of graphical or gesture-based secrets is usually easy to observe. In terms of usability, it was shown that most graphical authentication mechanisms are easy to use [41] as success rates are usually high. On the downside, the input often takes significantly more time compared to traditional passwords [32, 104]. This Section presents the most relevant examples of each category, discusses particular assets and drawbacks and the applicability to touchscreen mobile devices.

2.2.1 Recognizing Objects

Recognition-based systems are sometimes also referred to as cognometric [68, 104] or searchmetric [204] concepts. Such systems generally rely on identifying previously seen images [32, 104]. As the recognition of an item is usually easier than its recall [189, 235], such passwords are assumed to be particularly easy to remember. During enrollment, users memorize several pictures. While authenticating, these pictures are provided in line with a number of decoy images and users have to identify the previously memorized image set [32].

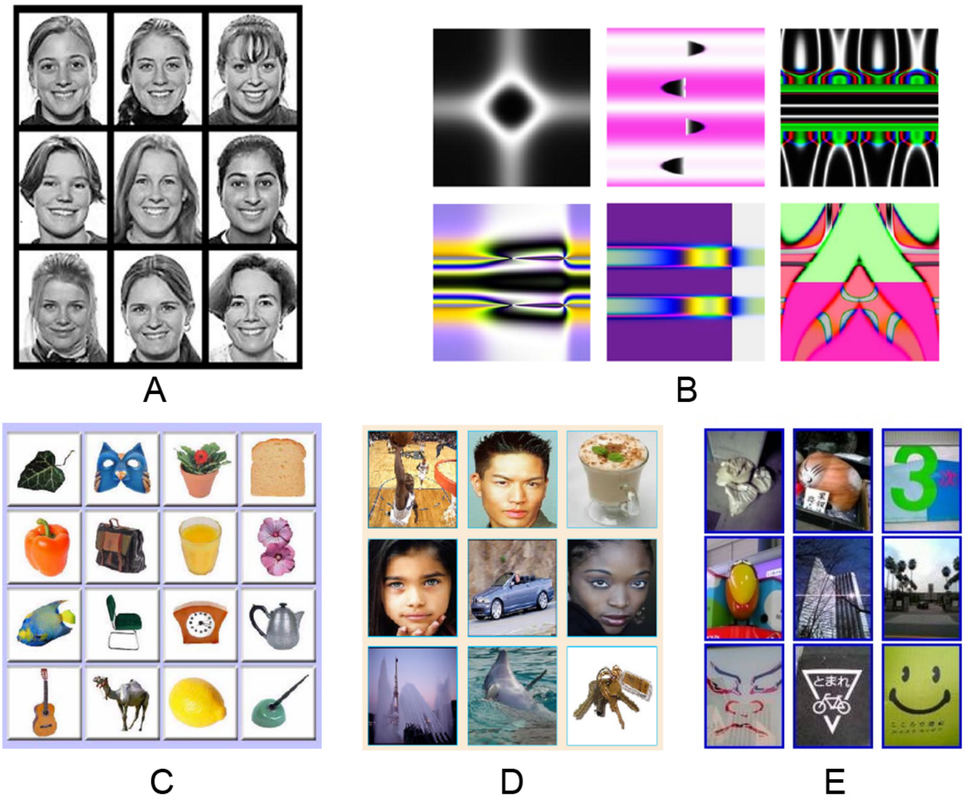


Figure 2.1: Representative recognition-based authentication concepts. (A) Passfaces [41], (B) Déjà Vu [77], (C) VIP [68], (D) Story [66] and (E) Awase-E [245].

Figure 2.1 illustrates five representative systems. Typical design factors, which are likely to have an impact on the usability and on the security of the resulting system, are *image type*, the *number of presented images* and the *relationship of the images* (e.g., similarity). While the image type influences the memorability of the passwords [41, 77, 245], a large number of decoy images significantly increases search times and therefore reduces the overall performance [68]. Similar images are harder to attack, but also harder to memorize [81]. The following examples will illustrate the effects of different design decisions in detail.

Representative Examples

Passfaces [41] which is commercially sold by the Passfaces Corporation[1] is the most researched recognition-based authentication system [204]. The Passfaces system is based on the recognition of human faces since they are assumed to be "[..] much easier to remember than passwords or PINS" [89]. During enrollment, the user selects four faces (see Figure 2.1 A). During authentication, the user needs to correctly identify all four faces, each within a grid of nine images. While the challenging sets are fixed, the faces are randomly positioned each round [41]. Brostoff and Sasse [41] claimed that Passfaces are significantly

[1] http://www.passfaces.com – last accessed: 2015/07/22

less error-prone than alphanumeric passwords. Valintine [256] indicated improved memorability. However, Davis et al. [66] revealed that the selection of Passfaces is very predictable. Their experiment showed significant effects of attraction, gender and race. Both genders selected female faces more often than male ones. Furthermore, attractive persons were selected more often and participants preferred people of their own race.

Dhamija and Perrig [77] propose Déjà Vu which is based on abstract random art images (see Figure 2.1 B). Since such pictures can be automatically generated, Déjà Vu does not require large image databases. A comparison to photographic pictures revealed that abstract images are harder to describe (which hampers password sharing) and that password selection is less predictable. A comparison to alphanumeric passwords showed that Déjà Vu is less error-prone, but authentication takes more time (< 30 seconds). De Angeli et al. [67, 68] designed the VIP system which provides "detailed, colorful and meaningful photos of objects" (see Figure 2.1 C). To prevent predictable password selection, images were randomly assigned to each user. A user study indicated that users prefer VIP to PIN, but no memorability benefits were found and the authentication process was significantly slower [68]. The authors tested different configurations using random and fixed token sets and concluded that a randomized presentation significantly decreases usability. Davis et al. [66] proposed to improve memorability and password selection by supporting story-based portfolios. Therefore, the user selects a sequence of unique images illustrating everyday objects to make a story (see Figure 2.1 D). Takada et al. [245] presented "Awase-E", a recognition-based scheme based on the user's personal photo collections (see Figure 2.1, D). They showed that using personal photos improves memorability, but at the same time makes guessing attacks more effective.

Assets and Drawbacks

Recognition-based approaches constantly outperform alphanumeric passwords in terms of memorability and success rates. This is especially true whenever meaningful objects or personal images are used. The biggest drawback according to usability is a high authentication time as multiple rounds of slow visual search tasks are usually required. This search time increases with more complex images and larger image sets [204]. While nameable and meaningful images increase usability, security might be downgraded through predictable password choice and password sharing. The use of similar and abstract images can increase security but makes the recall harder and further slows down the process [82].

While these concepts provide increased phishing protection as they present a predefined challenging set which is hard to fake [32, 104], secrets usually cannot be hashed. This is a significant drawback as secure sever-side storage is not possible and the information is usually available to anyone who gains access to the database [32]. In addition, the theoretical password space is usually small [32, 104, 113] which makes simple brute-force attacks efficient. Renaud and De Angeli [204] further note that observation attacks are often a problem, especially when touch screens are used [249].

Mobile Device Applicability

De Angeli et al. [67] argue that the "touch-screen is the ideal solution for the VIP paradigm", Dhamija and Perrig [77] recommend Déjà Vu especially for settings where text input is difficult. Indeed, some recognition-based authentication mechanisms were specifically designed for mobile devices (e.g., [122, 138, 245, 270]) and interacting with images on touchscreens is a common task. Dunphy et al. [80] analyzed the usability and the security of recognition-based systems on mobile devices in detail and found out that these concepts are indeed easy to use and actually less easy to observe than expected. However, long authentication times make these concepts infeasible. Dunphy et al. [80] note that "login durations of approximately 20 seconds are unattractive to many users". This should be especially true in the mobile context where device unlocks are usually performed many times a day.

2.2.2 Localizing Regions

Concepts which are based on the localization of specific regions within a given picture are generally referred to as cued-recall [32], click-based [57] or locimetric [104] systems. In such concepts a picture is given as an external cue to support the recall of specific click-points within the picture. The cues should only support the user and not the attacker [32]. During the enrollment, the user selects multiple regions within one or more pictures. When authenticating, the user needs to provide the right click-points, usually in a specific order [104]. Examples of such systems are given in Figure 2.2.

Typical design factors, which impact the usability and the security of the resulting system, are the *tolerance level* [277], the *number of pictures* [57] and the *visual properties* of the used image [78]. As humans are unable to accurately select single pixels of a given picture, the tolerance level is the main adjuster for the security and the usability [277]. Higher tolerance levels increase the ease of use while lower tolerance levels increase the theoretical password space. The number and visual properties of the images have significant effects on the practical security of the system in form of guessability [78] and observability [57]. The next Section will present example concepts to illustrate these factors.

Representative Examples

The oldest graphical password scheme proposed by Blonder in 1996 [35] is a locimetric concept. Passwords were based on predefined sequences of clicks in specific regions of an image [104]. The proposed interface is depicted in Figure 2.2 A. While Blonder's scheme was not further evaluated, its successor, the PassPoints authentication system was analyzed in detail [276]. PassPoints [276] is based on Blonder's idea but does not restrict the clickable area and supports a wider range of image types. Instead of providing predefined click areas, Wiedenbeck et al. [276] discretize the image to allow free selection. Furthermore, this mechanism supports securely hashed storage of the passwords. The authors performed a longitudinal user study using the interface illustrated in Figure 2.2 B. They found out that memorability and ease of use were similar to alphanumeric passwords, but authentication

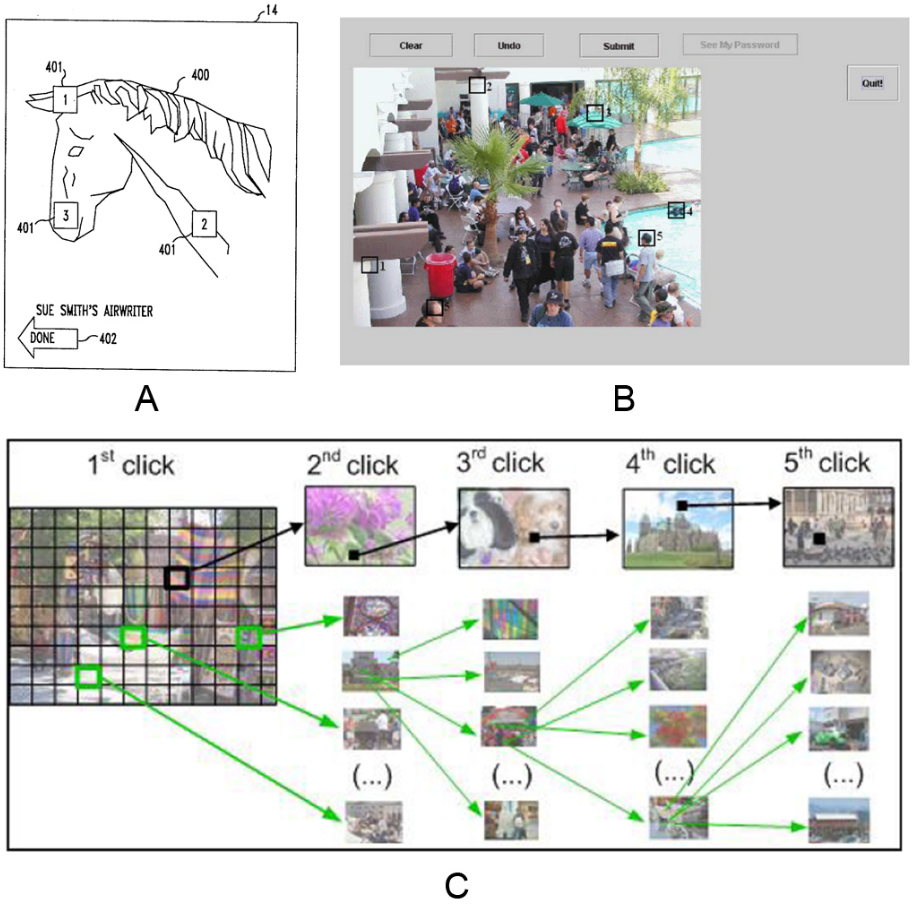


Figure 2.2: Representative cued-recall-based authentication concepts. (A) Blonder's patent [35], (B) PassPoints [276] and (C) Cued Click-Points [54].

took more time (19 seconds). In a second experiment, Wiedenbeck et al. [277] found that small tolerance values significantly reduce usability and memorability and thus a minimum tolerance of 14*x*14 pixels should be provided. In addition, a "hot spot" problem was discovered. This problem is based on the fact that humans do not click on all pixels of a given image with the same probability, but prefer specific attributes like objects, strong colors and high contrasts [78]. In a later experiment, Dirik et al. [78] modeled the user choice for a given image based on such visual properties. They were able to correctly predict up to 80% of the users' click positions. Chiasson et al. [57] proposed the Cued Click-Point concept [57] where users select one single click point on a sequence of five distinct images. The mechanism, which is illustrated in Figure 2.2 C, provides implicit feedback to the user as the presented images are determined by the click position in the current image. On the other hand, attackers cannot gain any information. A lab study revealed that the one to one relationship between images and click-points is easier to use (96% success rate) and at the same time more secure. On average, users needed seven seconds to successfully authenticate. The good usability was confirmed in a field study [53].

Assets and Drawbacks

Cued-recall systems provide good memorability and high success rates. The authentication process is faster compared to recognition-based concepts but remains slower than alphanumeric input. Furthermore, users seem to cope better with multiple accounts when using click-based passwords, as they select a more diverse set of passwords compared to the text-based scenario [55]. Finally, locimetric passwords support secure (hashed) storage [276].

On the downside, the input of click-based passwords is usually easy to observe [104]. This is especially true when the concept is based on one single image (e.g., PassPoints). In addition, users follow distinct patterns when creating click-based passwords [54]. The focus on specific hot spots (e.g., objects) within the image and the use of simple geometric patterns make cued-recall systems vulnerable to automatic dictionary attacks [259,296]. Renaud and De Angeli [204] conclude that "[..] any picture has a limited number of distinct objects [..] and this makes locimetric systems untenable".

Mobile Device Applicability

Current mobile devices provide relatively small touch-based interfaces [214]. Therefore, precisely selecting small regions of a picture becomes a usability problem [52, 146]. Schaub et al. [214] developed a mobile version of the Cued Click-Points concept and reported that the tolerance squares had to be increased to 98*x*98 pixels which resulted in a significantly reduced theoretical password space. Suo [243] tested click-based passwords on iPads and achieved success rates of 80% using a 30 pixels tolerance. Stobert et al. [239] analyzed the interplay of image size and the number of click-points. They found no usability drawbacks in using more click-points with smaller pictures and therefore suggest using this configuration on mobile devices. However, the reduced resolution in combination with biased password selection generally questions the applicability of such concepts to mobile devices.

2.2.3 Recalling Shapes and Gestures

Recall-based graphical password systems are also referred to as drawmetric schemes [204]. Such systems are based on the recall of previously memorized information without additional cues [32]. Most of these systems are based on reproducing graphical shapes on a grid-based canvas (e.g., [103, 140, 248]). Figure 2.3 illustrates prominent representatives. Simple shapes can also be seen as two-dimensional touch gestures [227]. More recently published concepts indeed forgo visual feedback and should therefore be categorized as gesture-based concepts (e.g., [208, 227]). Please note that apart from touch gestures, other three-dimensional gesture-based systems were proposed. However, such schemes which utilize in-air hand gestures (e.g., [59, 111, 165, 176]) or full body movements (e.g., [123, 173]) are out of scope of this work and will not be discussed.

Important design factors, which influence the usability and the security, are the *matching tolerance* [103], the *degree-of-freedom* of the input [248] and the provided *visual feedback* [260]. As precisely reproducing previously entered shapes is very difficult, a relaxed matching mechanism increases usability. On the other hand, the theoretical password space is reduced. The degree-of-freedom determines if shapes are freely drawn or if the drawing is bounded to a grid. Finally, visual feedback can ease the reproduction of shapes but at the same time increases observability. The different factors are illustrated in the next Section.

Representative Examples

In 1999, Jermyn et al. [140] published the first drawmetric password system with the goal to achieve better security on personal digital assistants (PDAs). The proposed system which was called Draw-a-Secret (DAS) is depicted in Figure 2.3 A. Draw-a-Secret passwords consist of an arbitrary number of strokes which are drawn on a $5x5$ grid. The resulting password is mapped to a sequence of grid coordinates and allows hashing. While the theoretical password space of DAS is comparable to alphanumerical passwords (2^{58}) [140], it was shown that user choice is predictable and DAS is vulnerable to dictionary attacks [187, 252]. User-selected drawings are usually short [252] and placed in the center of the grid [187]. In addition, users tend to choose symmetric shapes [187]. Varenhorst et al. [260] presented Passdoodles which allows free-form drawing. In contrast to DAS, drawings are based on a canvas without visible grid and can be composed of different colors. Qualitative Draw-A-Secret (QDAS) [164] encodes drawings based on qualitative direction changes instead of relying on grid-cells. As a consequence, users are not required to memorize concrete grids but recall the starting cell and the order of direction changes.

Gao et al. [103] proposed another modification of DAS, called YAGP (Yet Another Graphical Password). YAGP provides a larger grid ($48x64$) and allows position-free and size-independent drawings (see Figure 2.3 C). First results indicated that users were able to fend off observation attacks by drawing smaller versions of their passwords at covered positions. On the downside, the matching algorithm of YAGP does not support hashed storage of secrets. Finally, PassShapes [275] combines the ideas of Passdoodles, QDAS and YAGP. The PassShape interface does not provide a grid as the algorithm allows to draw on any position of the canvas. As indicated in Figure 2.3 D, PassShapes are based on a predefined set of eight different directions. The evaluation showed that this aspect simplified the input and improved memorability. At the same time, the theoretical password space was scaled down.

Tao and Adams [248] presented Pass-Go. In contrast to earlier concepts which were based grid cells, Pass-Go utilized the intersecting lines (see Figure 2.3 E). When a sensitive area at an intersection is touched, a predefined shape appears. For single touches, a dot is displayed, continuous touches are indicated by lines. As a consequence, the visualization of Pass-Go passwords is unaffected from (small) trace variations and the indicators can be optimized for good perceptibility. The concept furthermore supports switching off the indicators to increase security. A three-month field study revealed an average password composition of five strokes and two dots [32, 248]. The semantic analysis of the drawings revealed predictable selection patterns. Users based their passwords on alphanumeric forms or used symmetric shapes. In addition, many users started their pattern in the upper left corner and finished in the lower right. While only three percent of the logins were performed without visual indicators, 76% of them were successful.

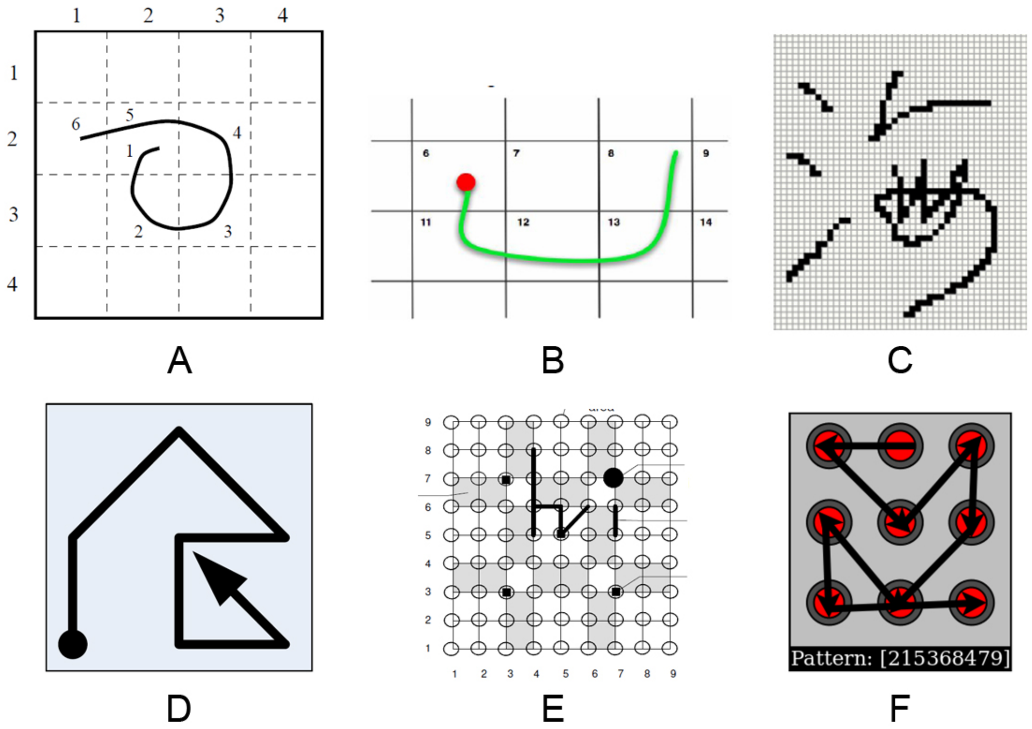


Figure 2.3: Representative recall-based authentication concepts. (A) Draw-a-Secret (DAS) [140], (B) Qualitative Draw-A-Secret (QDAS) [164], (C) Yet Another Graphical Password Strategy (YAGP) [103], (D) PassShapes [275], (E) Pass-Go [248] and (F) Android unlock pattern [18].

In recent years, the Android unlock pattern became the first graphical password scheme which was widely accepted [104]. The authentication mechanism which is available on all mobile devices with Google's Android operating system is a simplified version of Pass-Go [104]. As seen in Figure 2.3 F, Android unlock patterns describe simple continuous shapes which are entered on a 3*x*3 grid. The input is usually visualized through line and point indicators. A valid pattern must comprise a minimum of four grid cells. As each point can only be activated once and stays activated thereafter, the maximum length of a pattern is nine. In addition, points which are orthogonal neighbors cannot be bypassed without activation [18]. Due to these restrictions, the theoretical password space of the Android pattern unlock is 389,112 [18]. Uellenbreck et al. [254] analyzed the selection strategies of 113 users and found that most users start their patterns on the left side of the matrix and end on the right side. In addition, the upper left corner was the most used starting point. The results indicated that Android patterns are vulnerable to dictionary attacks. Aviv et al. [18] revealed the feasibility of smudge attacks. These attacks exploit oily traces (smudge) which the user's fingers leave on the touchscreen. In a user study, 68% of the unlock patterns were exposed through such an attack. Andriotis et al. [11] confirmed the predictability of user-selected Android patterns and claim that combining this knowledge with physical attacks based on smudge or observation makes unlock patterns very vulnerable. Due to the reduced degree-of-freedom, Android unlock patterns rather represent gestures than drawings. This is especially true when visual indicators are switched off.

Finally, some systems generally forgo visual feedback. Sherman et al. [227] proposed to use free-form multi touch gestures on mobile devices without visual indicators to prevent observation attacks. Niu and Chen [192] base the password on simple taps and argue that this gesture class is easy to reproduce for the user, but hard to do for the attacker. Sae-Bae et al. [208] evaluated five-finger gestures on touchscreens and additionally analyzed the user's input characteristics. A similar concept was proposed by Sun et al. [242]. Azenkot et al. [20] presented PassChords, a tap based authentication system specifically designed for blind users.

Assets and Drawbacks

Graphical recall-based authentication systems fundamentally differ from the other two categories. As the authentication is based on un-cued recall, these concepts have typically no need for external images. This aspect increases the theoretical password space and makes most drawmetric concepts more secure against brute force attacks [104, 113]. In addition, most concepts allow hashed storage of passwords [32]. Beside the pictorial superiority effect [226, 235, 236], the repeated reproduction of the same shape or gesture can stimulate the muscle memory [3, 94, 221] and therefore further improve memorability [204]. Furthermore, such concepts allow faster authentication than other graphical password systems [14]. This is especially true, when the input is based on simple shapes and gestures [275]. Finally, gesture-based authentication mechanisms allow eyes-free input and therefore support visually impaired users [20].

On the downside, drawmetric concepts suffer from similar problems like alphanumeric passwords. Password selection is often predictable which makes the secrets vulnerable to dictionary attacks [146,197,254] and users can easily "write down" or describe their passwords as drawing them is part of the concept [32]. Furthermore, pure recall concepts cannot provide improved phishing protection like the other two schemes [104]. As precisely reproducing the same shapes or gestures is a difficult task [146,204,248], the complexity of feasible secrets is restricted. In addition, this class of concepts is vulnerable to physical attacks [18, 104, 214]. Firstly, the direct input of simple drawings or gestures is usually easy to observe [104, 214]. Secondly, when gestures are entered using a touchscreen, smudge attacks become a serious problem [11, 18]. This is especially true when the input is based on a fixed grid.

Mobile Device Applicability

The literature review reveals that all drawmetric systems were specifically designed for the application on mobile devices. However, earlier concepts assumed that users interact with mobile devices using a stylus and therefore required detailed drawings. As current mobile devices are providing touch interfaces and finger-based interaction has become common, the reproduction of complex drawings is more difficult and significantly slower. This was confirmed by Chiang and Chiasson [52] who tested Draw-a-Secret on smartphones. Schaub et al. [214] adapted Pass-Go to mobile device and reported that the grid resolution had to be reduced to 5*x*5 and the sensitive areas had to be increased. These modifications made Pass-Go more similar to Android's unlock patterns. Indeed, usability-optimized drawmetric systems like Android patterns which are focusing on simple gestures and not on detailed drawings seem well suited for mobile devices. However, current solutions are prone to several threats like smudge attacks and observation. Therefore, the practical security must be improved to make gesture-based schemes a promising alternative for PINs and alphanumeric passwords on mobile devices.

2.2.4 Hybrid Approaches

Hybrid concepts comprise two or more aspects of recognition-based, cued-recall or pure recall schemes [104]. According to the used combination, such systems provide similar usability and security characteristics like single concepts [104]. Therefore, this Section merely illustrates some examples, but does not provide a detailed discussion of usability and security features. Some examples of hybrid authentication concepts are depicted in Figure 2.4.

Alsulaiman and Saddik [5] proposed the 3D Graphical password, an authentication scheme which is based on the interaction with virtual objects in a three-dimensional environment. It is one of the first hybrid concepts where input can involve mouse gestures (recall), switching of virtual lights (cued-recall) or recognizing objects. Thus, the concept comprises all three mechanisms and allows the user to freely combine these aspects in a password. On the downside, navigating through virtual worlds is a time-consuming task which requires the undivided attention of the user. Therefore, even if 3D passwords were specifically designed for mobile devices, they seem to be unhandy for daily use. Citty and Hutchings [61] presented TAPI which combines cued-recall and recognition-based aspects. TAPI presents a sequence of four challenging sets. As illustrated in Figure 2.4 A, each set consists of 16 images which are segmented into four parts. To authenticate, the user has to recognize the right objects and then indicate the correct segment within each image. TAPI was found to be more secure against observation attacks with authentication times from three to eleven seconds. Yuxin Meng [179] combined object recognition with gesture recall. According to Meng's concept, a user first recognizes and selects four images in a specific order. In a second step, the user performs a gesture on one of the selected images. Preliminary results indicated high success rates but login times over 13 seconds. Gao et al. [105] presented a recognition-based scheme which utilizes drawmetric elements as input method. To authenticate, the user draws a line through a predefined sequence of images resulting in a shape. As images are presented in random order, the resulting shape changes at each login. Since the drawn line activates both, secret images and decoy images, the system is resistant against observation (see Section 2.3). However, as for most recognition-based concepts, passwords need to be stored unencrypted.

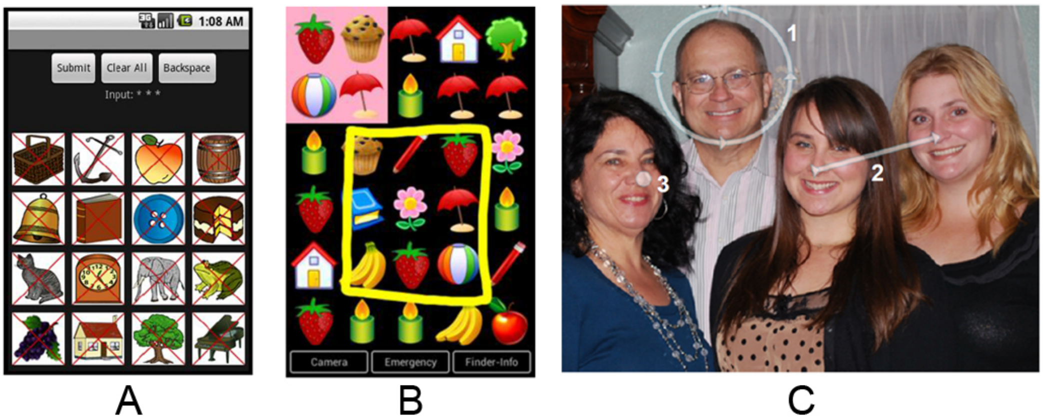


Figure 2.4: Examples of hybrid graphical authentication concepts. (A) TAPI [61], (B) Gesture-Puzzle [216] and (C) Picture Gesture Authentication (PGA) [295].

Schlöglhofer and Sametinger [216] proposed GesturePuzzle, a cued-recall mechanism for simple gestures. During enrollment, the user selects one or more objects and combines the set of objects with a specific gestures. During authentication, the displayed objects indicate the requested gesture. Figure 2.4 B illustrates the mechanism: the icon set is presented in the highlighted area. The user would recognize the specific set and perform a rectangle gesture (which was defined during enrollment). With Windows 8, Microsoft widely deployed a hybrid concept called Picture Gesture Authentication (PGA) [295]. PGA combines the original idea of PassPoints with discrete gestures. Instead of tapping, users perform gestures on specific regions of a background image. As indicated in Figure 2.4 C, the gesture set comprises circles, straight lines and taps. While PGA provides a large theoretical password space, Zhao et al. [295] revealed that PGA has the same problem like traditional locimetric systems. Users tend to perform their gestures on specific hotspots of the image which makes the system vulnerable to dictionary attacks. At the same time, the authors claim that the system provides more usability benefits than most other graphical password schemes.

2.3 Improving Security through Graphical and Gesture-based Concepts

Section 2.2 indicated that graphical mechanisms indeed have advantages over traditional text-based passwords. For example, graphical password systems are harder to break with traditional attacks like social engineering [104] and are often easier to remember [32]. At the same time, the evaluation revealed that simply using graphics instead of letters and symbols does not solve the password problem [64, 90]. User choice remains predictable for many concepts and observation attacks are a significant threat [104]. In addition, the use of such concepts on touch-based devices can open new security holes like smudge attacks [19].

To reduce these problems, researchers started to follow HCI principles to find more secure but yet usable interaction methods and presentation styles [104]. The following Section presents authentication concepts which where designed to reduce the threats of guessability, smudge attacks and observability. We illustrate that increasing the practical security is a relatively easy task, but becomes exceptionally hard when usability benefits shall be preserved. Beside the presented concepts, other approaches exist which for example exploit biometric features to provide additional security (e.g., [12, 69, 222]). As pointed out in Chapter 1, biometric solutions are out of scope of this work and will not be discussed.

2.3.1 Resistance against Guessing Attacks

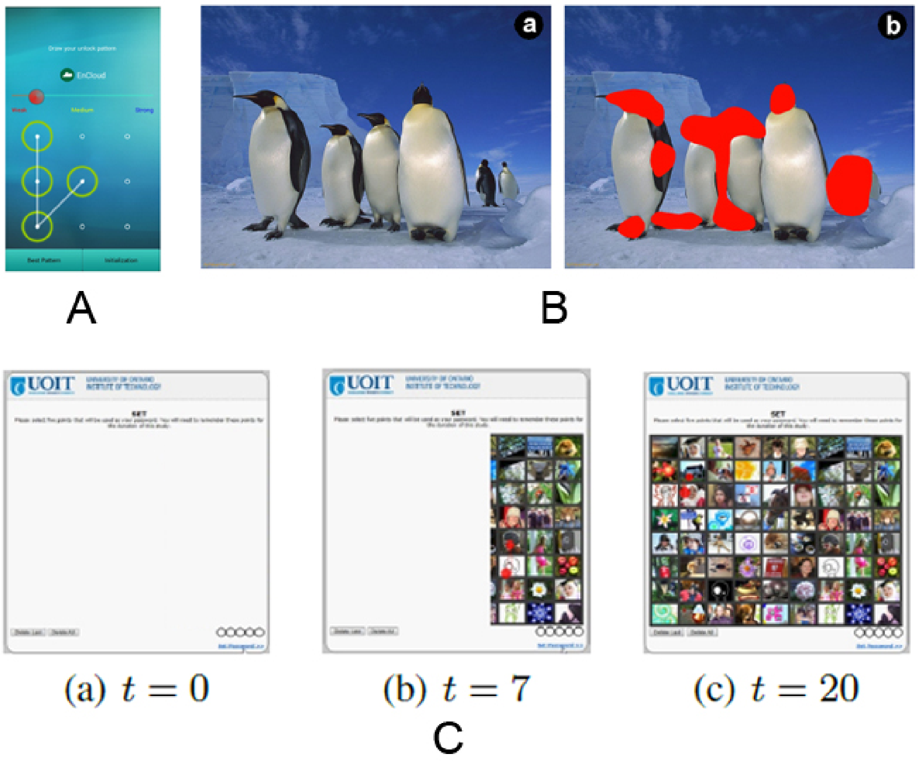
Section 2.2 indicated that all presented graphical and gesture-based password systems are vulnerable to guessing attacks. Most recognition-based approaches provide a small password space and are vulnerable to brute force attacks. Locimetric schemes suffer from hot spot problems and allow for fully automatic guessing attacks. Drawmetric schemes are vulnerable to dictionary attacks as users choose predictable shapes and gestures. Though these problems are known for several years, related work concerning the secure selection of graphical or gesture-based secrets is relatively sparse. Conservative solutions imply system-generated password assignments (e.g., [67, 89]) and password policies [197]. However, the lessons learned from text-based passwords (see Section 2.1) indicate that such restrictive countermeasures reduce memorability and introduce insecure behavior. As a consequence, we are interested in concepts which exploit the possibilities of user-centered interaction design to positively influence selection behavior and to support users in creating better secrets.

This Section presents such approaches and categorizes the mechanisms based on the Persuasive Authentication Framework by Forget et al. [99]. The Persuasive Authentication Framework provides five principles which can be used to positively influence password selection. The presented concepts are categorized based on four of these principles, namely monitoring, personalization, simplification and conditioning.

Monitoring the Input

The monitoring principle is based on the idea that users who are being observed are more likely to behave in the desired way [99]. A special type of monitoring is the self-monitoring where users get feedback to their behavior. Password meters which are already known from alphanumeric passwords are one example of this category as the user gets feedback on the strength of the password input. Several projects have considered this mechanism for gesture-based authentication mechanisms, more precisely for Android unlock patterns. As depicted in Figure 2.5 A, the visual appearance of such concepts matches the design of alphanumeric password meters. However, the assessment of pattern strength differs between the concepts.

Sun et al. [241] calculate pattern strength based on the number of straight lines, the number of dots (length), intersections and overlapping dots. Andriotis et al. [10] base the pattern strength on length, direction changes, knight moves and overlaps. Finally, Song et al. [232]



(a) $t = 0$ (b) $t = 7$ (c) $t = 20$

Figure 2.5: Designs to improve graphical password selection. (A) Pattern Meter [232], (B) Saliency Masks [45] and (C) the Presentation Effect [251].

proposed to evaluate pattern length, the number of non-repeated sub-patterns and intersections. The relative weights of such pattern characteristics have not been evaluated. However, it was shown that users react by selecting more complex gestures [10,229,232,241]. Siadati and Memon [229] simulated guessing attacks on traditional unlock patterns and patterns selected with a pattern meter. They report that 50% of the traditional patterns could be guessed within 1000 guesses, while only 22.6% of the passwords of the meter condition were revealed. Song et al. [232] attacked 101 patterns of both groups and required 16 guesses to expose 10% of the traditional patterns and 48 guesses to reveal the same number under the pattern meter condition. Unfortunately, the user studies also revealed that the choice of the starting point of the gesture is not influenced by pattern meters. Apart from these concepts concerning unlock patterns, the monitoring principle has not been proposed for graphical password systems.

Personalizing the Interface

The personalization principle is based on the idea that personalized information is more persuasive than generic information. Forget et al. [99] distinguish this principle in suggestion and tailoring. Personalized interfaces have been proposed for the enrollment of drawmetric and locimetric passwords.

Dunphy et al. [83] proposed Background Draw a Secret (BDAS). BDAS is a modification of Draw a Secret (DAS) which allows users to select personal background images. The authors compared DAS and BDAS in a lab-based user study and found out that users selected more complex and less symmetric patterns. In addition, the selected secrets were significantly longer and less centered. A recall test after one week indicated that background images additionally improved memorability. On the downside, the authors note that the images might introduce hot spot problems known from locimetric approaches and some participants claimed that the colored background was distracting. Nevertheless, the BDAS project indicates that simple changes in the user interface (like personal background images) significantly influence password security. Por et al. [202] confirmed these positive effects with a version of Pass-Go which also utilized personal background images.

Thorpe et al. [251] evaluated presentation effects on graphical password selection using Pass-Points. For this purpose, they modified the presentation of the background image. Instead of displaying the whole image at once, they opened a white curtain either from the left or from the right side of the screen. The effect took 20 seconds and the image was revealed at constant rate (see Figure 2.5 C). The evaluation of the system revealed a significant impact on the click-point selection. Users choose the first click point at the side of the screen which was first revealed. The authors conclude that such simple modifications can result in a significantly different password distribution and note that the specific presentation used during enrollment should be unknown to the attacker. Siadati and Memon [229] proposed a similar effect for Android unlock patterns. During enrollment, one random point of the $3x3$ matrix starts blinking to nudge users to select this specific point first. The effect resulted in nearly equally distributed starting points. However, pattern composition remained predicable.

Chiasson et al. [54] presented a modification of the Cued Click-Points (CCP) scheme called Persuasive Cued Click-Points (PCCP). The evaluation of CCP had already indicated that the sequential ordering of multiple pictures can reduce the hotspot problem. PCCP was designed to further eliminate hotspots by persuading users to select more random click-points [56]. During enrollment the system highlights a small rectangle which is randomly positioned on the background image. Users are required to select a click-point within the given rectangle. To ease the selection, the rectangle can be shuffled. The analysis of the concept revealed that users are indeed nudged to select more random click-points and thus the practical password space can be significantly increased. Since the active selection area is reduced and shuffling the rectangle implies additional effort, the system also comprises aspects of simplification and conditioning [99]. These principles are discussed in the next Section.

Simplifying the Selection

Forget et al. [99] distinguish the simplification principle in tunneling and reduction effects. The main goal of such strategies is to reduce the complexity and the number of required interactions. This can be done by suggesting secrets or by reducing the number of available options. In contrast to personalized strategies like PCCP [56] which provide personalized suggestions for each user, concepts of this category display the same set of options to all users.

Bulling et al. [45] proposed a simplification approach for PassPoints. The concept is based on a computational model for visual attention and masks areas of an image which are most likely selected by the user (see Figure 2.5 B). However, these saliency masks are only present during enrollment. During authentication, the whole image is displayed. A user study indicated that saliency masks indeed reduce hotspot selections and increase the practical security of PassPoints. Uellenbeck et al. [254] proposed to change the layout of the Android pattern unlock to reduce the predictability of user-selected secrets. One idea was to remove the upper left cell as this cell is most often chosen as a starting point. However, the change did not have the desired impact as users simply chose the point next to the removed cell. Other proposed layouts were quasi-random or based on a circular design. While these layouts effectively prevented selection patterns known from the matrix layout, they introduced other predictable selection behavior. This shows that simply changing from one layout to another is not an effective solution and interface changes need to be thoroughly thought through.

Bicakci et al. [31] proposed to solve the hot spot problem of click-based password schemes by assigning random passwords. Therefore, the background image provides a set of well distinguishable icons. During enrollment, the system called GPIS randomly generates a set of icons (click-points) and suggests the secret to the user. If the user is not satisfied with the generated password, she can ask the system to generate a new one. This procedure can be infinitely repeated. The evaluation showed that the system effectively eliminates the hotspot problem. The concept is similar to the Persuasive Cued Click-Points (PCCP) scheme but further restricts user choice [56]. Finally, both concepts may exploit the principle of conditioning as not accepting the suggested viewport (PCCP) or icon set (GPIS) results in additional effort for the user. According to the conditioning principle, desired behavior shall be reinforced. As accepting the suggested click-points results in a faster enrollment, both systems reinforce the compliance with the recommendations.

2.3.2 Resistance against Smudge Attacks

The design of authentication systems which are particularly resistant against smudge attacks is a relatively new discipline as the susceptibility of grid-based passwords to smudge attacks was just recently revealed [19]. Searchmetric concepts (e.g., Passfaces [41]) usually rely on a randomized presentation of challenging sets, and thus provide inherent smudge attack security. In contrast, the static input of locimetric and drawmetric concepts allows the interpretation of smudge traces to derive the entered secret. Up to now, there are only three publications specifically focusing on secure designs against smudge attacks [4, 161, 218]. The presented concepts prevent smudge attacks through the introduction of randomization and additional tasks.

Airowaily and Alrubaian [4] presented WhisperCore which is based on the generic idea of blurring smudge traces. For this purpose, the user is required to wipe a presented figure as a final step before the smartphone is unlocked (see Figure 2.6 A). The mechanism works independently from the used authentication mechanism and can be added to any vulnerable

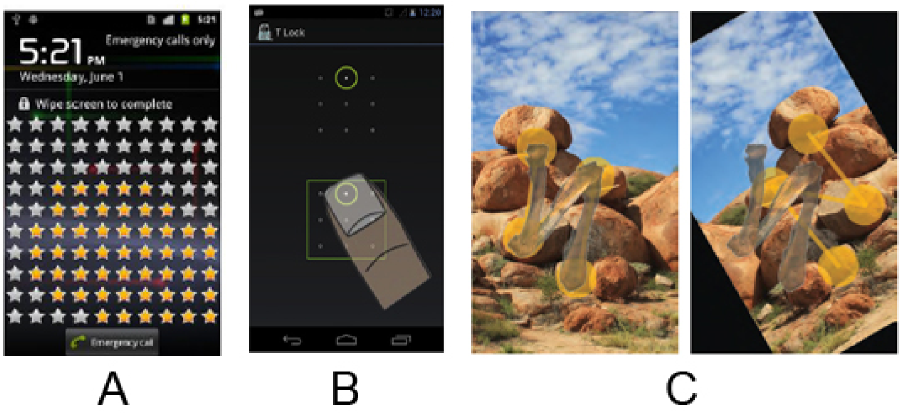


Figure 2.6: Smudge attack resistant concepts. (A) WhisperCore [4], (B) TinyLock [161] and (C) SmudgeSafe [218].

system. The authors did not report any evaluation data. However, it can be assumed that the introduction of an additional wiping task will annoy most users. Kwon and Na [161] proposed TinyLock, a smudge attack resistant concept for unlock patterns. As indicated in Figure 2.6 B, the concept introduces a minimized $3x3$ grid and an additional wiping task. Unlock patterns are entered using the minimized grid. After the pattern is entered, a virtual wheel needs to be rotated at the same position to blur smudge traces. TinyLock was compared to the original unlock pattern scheme. Even if TinyLock slowed down the authentication process, it was equally easy to use as the original scheme. The security analysis showed that TinyLock was remarkably secure against smudge attacks as none of the entered patterns was exposed. Schneegass et al. [218] presented SmudgeSafe, a drawmetric system based on background images. Similar to PGA [295], the user performs a gesture on specific regions of the background image to authenticate [202]. To increase the resistance to smudge attacks, geometric image transformations are randomly applied during authentication (see Figure 2.6 C). The authors tested a transformation set consisting of translation, rotation, scaling, shearing, and flipping. The security analysis revealed that 5-30% (avg. 12.7%) of the smudge traces could be interpreted depending on the used transformation. The usability analysis revealed an average authentication time of 3.6 seconds and an error rate of 25%.

2.3.3 Resistance against Observation Attacks

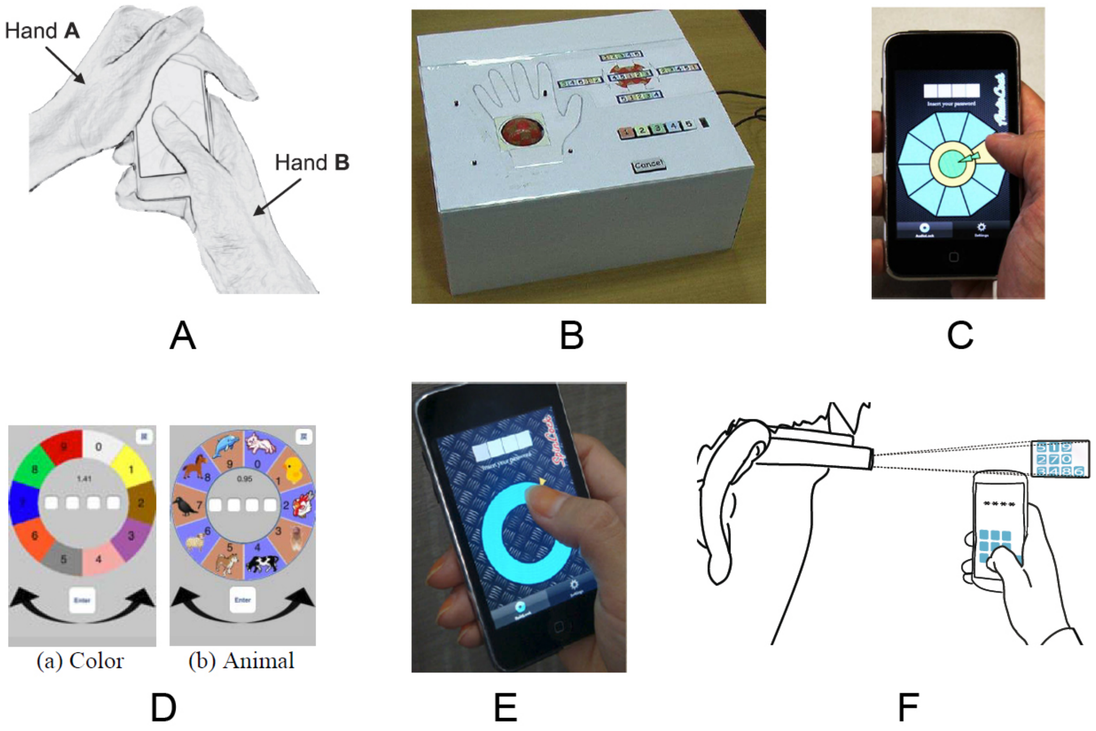
Observation attacks, also called shoulder surfing attacks, can be distinguished into cognitive attacks and video-based observations [276]. While the use of technical equipment like video cameras allows a detailed analysis of the observed input, cognitive attacks are restricted by the attackers short-term memory. In general, all knowledge-based authentication mechanisms are prone to such attacks [249]. However, graphical and gesture-based password systems are often exceptionally easy to observe as they are based on the direct interaction with graphical elements [104,249]. While the entry of alphanumeric passwords can be somewhat disguised using asterisk-based password fields, the input of a touch gesture is harder to shield [292]. The same is true for the selection of click-points [276] or pass-icons [81]. As a consequence, the development of graphical and gesture-based authentication mechanisms which are resistant to shoulder surfing became an active research area [279].

The design of hard-to-observe authentication mechanisms nicely illustrates the interplay of usability and security features. Proposed concepts often rely on complex selection tasks (e.g., [160, 246, 278]), cognitive detours (e.g., [112, 130, 147]) or the recall of additional secrets (e.g., [72,244,283]). While these strategies make the input indeed harder to observe, they also tend to increase the mental load and input times. Other concepts require additional hardware or increased computational power (e.g., [100, 281]). The discussed concepts are classified according to the used strategy. We distinguish three different approaches: The first approach establishes invisible communication channels. The second class is based on the indirect selection of tokens or on multiplexed input. The last concept type applies distraction strategies or visually overloads the graphical user interface. The following explains the three strategies in detail and discusses their usability and security.

Invisible Communication Channels

Perhaps the most obvious solution for the shoulder surfing problem is making (parts of) the authentication process invisible. Such invisible channels are established through nudging users to shield their display, through multi-modal interaction and by exploiting micro gestures like touch pressure or eye movements. Kim et al. [151] proposed ShieldPIN to protect the PIN entry on tabletops. To authenticate, the user has to perform a shield gesture in a specific area on the screen. As soon as the shield gesture is performed, the keypad appears and the PIN can be entered. Qiang Yan [288] proposed CoverPad, a similar approach for mobile devices (see Figure 2.7 A). CoverPad additionally delivers hidden transformation messages (e.g., “add 3 to every digit”) whenever the shielding gesture is recognized. As the user transforms her input according to the presented message, the observable interaction with the keypad does not reveal the memorized secret.

Sasamoto et al. [211] were the first to present a multi-modal authentication approach relying on sensory cues. The searchmetric system, called Undercover, is based on a portfolio of five previously memorized pictures. In several rounds, the user is asked to identify her picture within a presented challenge set. The challenge set consists of four pictures and a “none” option. To secure this visual channel, the user establishes a secret channel by placing the left hand on a machine which delivers sensory cues. The final prototype is depicted in Figure 2.7 B. According to the sensed cue, the user gives the right or a wrong answer (binary cue) or transforms the answer. The authors tested contact-based, friction-based and tactile cues and claimed that Undercover is secure against repeated observation and video-attacks assuming that hand movements and motor noises cannot be interpreted. De Luca et al. [73] presented a similar concept for the secure input of alphanumeric secrets on public terminals. Similar to Undercover, Vibrapass delivers binary haptic cues to indicate fake input (so called lies). The authors proposed to exploit the vibration function of the user’s mobile device to establish the secret channel.



(a) Color (b) Animal

Figure 2.7: Observation resistant concepts based on a secret channel. (A) CoverPad [288], (B) Undercover [211], (C) PhoneLock [27], (D) Audio Instructions for Multisensory Authentication [120], (E) Spinlock [28] and (F) GlassUnlock [281]

While Undercover and Vibrapass were designed to protect the input on public terminals, multi-modal concepts were also proposed for mobile devices. Bianchi et al. [29] designed PhoneLock [27] which allows secure PIN-entry on smartphones. As illustrated in Figure 2.7 C, the system provides a graphical wheel which is composed of ten equally sized targets, each representing a digit (0-9). Although the orientation of the wheel is randomized, the internal order of the segments stays the same. As soon as the user touches one segment, one out of ten specific cues is delivered indicating the underlying digit. The authors tested haptic and audio cues and reported average authentication times between 12 (audio) and 28 (haptic) seconds depending on the modality. Hasegawa et al. [120] tested different audio cues on a similar wheel-based input system. Instead of simple peeps, the authors proposed to use animal sounds, colors or clock-based instructions. As depicted in Figure 2.7 D, the graphical wheel is illustrated according to the given cues. For example, when animal sounds are delivered, images of the corresponding animals are provided (see Figure 2.7 D(a)). The evaluation indicated that concrete sounds and images are less error prone and allow faster authentication times than abstract cues (e.g., peeps). Kuribara et al. [160] published VibraInput which is based on a reduced set of distinct vibration patterns. Instead of ten cues, VibraInput provides only four distinguishable patterns. Each digit is entered in two distinct steps using a graphical wheel. The entered digit is derived by the intersection of both steps.

To further reduce the effort of recognizing and interpreting distinct cues, Bianchi et al. [28, 29] proposed a set of authentication concepts which are based on counting simple cues. Spinlock is one example of this class. Similar to the prior concepts, Spinlock is also based on a graphical wheel (see Figure 2.7 E). As soon as the user starts rotating the wheel, the system delivers single cues (haptic or audio). After the intended number of cues has been received, the user stops the input and the number of cues is translated into a single digit and sent to the system. The evaluation of such single-cue systems indicated that counting is faster and less error prone [29]. Instead of relying on haptic or audio cues, Winkler et al. [281] presented GlassUnlock which secretly delivers visual cues through a private near-eye display (e.g., smart glasses). For example, the user would see the current mapping of a randomized PIN-pad through her smart glasses while the input would be performed on the empty buttons of the smartphone (see Figure 2.7 F). The authors claimed that authentication times were comparable fast to standard PIN-entry. The downside of the concept is the requirement of additional hardware.

In addition to shielding gestures and multi-modal interaction, the use of micro gestures has been proposed to protect authentication from observation attacks. TinyLock [161] which was discussed in connection with smudge attacks (Section 2.3.2) is one representative of this concept class. Kim et al. [151] proposed Pressure-Grid which allows the entry of graphical and numeric passwords. To authenticate, the user places three fingers on distinct areas of a touchscreen and presses the fingers to select specific items of a grid. The evaluation showed that pressure-based input is significantly harder to observe. Malek et al. [169] proposed to use pressure-based input for grid-based gestures. A performed lab study revealed only limited observation resistance. While most participants included pressure into their gestures, the qualitative feedback indicated that using the additional pressure information was annoying for some users. Gaze-input can be seen as another class of micro gestures as the selection of screen items requires very small eye movements. Weiss and De Luca [275] evaluated the input of PassShapes based on relative eye-gestures and reported that this input method slows down the authentication process but at the same time increases observation resistance. In a simulated video-based observation attack, 100% of the traditional PassShapes were exposed, but only 55% of the secrets entered with eye-gestures could be identified. Forget et al. [100] proposed Cued Gaze-Points, a version of PassPoints relying on eye-gaze input. To authenticate, the user looks at the desired point of the image and presses the space bar to confirm. Arianezhad et al. [15] applied eye-gaze input to grid-based graphical password schemes. The latter two concepts are significantly harder to observe, since the observed eye movements need to be mapped to specific regions on the display.

Indirect Input and Multiplexed Secrets

Authentication systems of this category are based on the indirect selection of secret tokens. Therefore, the user either touches an area distant from the actual secret or selects a larger set of tokens to disguise the actual secret. Roth et al. [207] presented the first representatives of this class, referred to as cognitive trapdoor games. The systems allowed shoulder surfing resistant PIN-entry. As indicated in Figure 2.8 A, the system provided a PIN pad, where half of the digits were colored in black and the other half was colored in white. To authenticate, the user repeatedly indicates the set (color) which comprises the intended digit by pressing one of the buttons labeled “black” and “white”. The system calculates the intersection of the subsets and derives the entered digit. As 16 challenges are required to enter a 4-digit PIN, authentication times are high (> 20 seconds). Takada [244] proposed to augment digits with shapes. According to the FakePointer concept, the user memorizes her PIN and one distinct shape per digit. During authentication different shapes are randomly assigned to the ten buttons of the PIN pad (see Figure 2.8 B). To authenticate, the user shifts the numbers using the left and right keys of the keyboard. As soon as the intended digit matches the memorized shape, the process is completed. De Luca et al. [72] added color information to each digit to allow the indirect input of PINs. The system, called ColorPIN, displays each digit with three differently colored letters. To enter a digit, the user presses the respective letter that matches the previously memorized color. In the example of Figure 2.8 C, the user would have memorized “4, black” and would therefore press the “L” on the keyboard. The main advantage of ColorPIN is the one-to-one relationship between user input and the length of the entered PIN. Like FakePointers, the downside is that the user has to memorize one additional secret per digit. A performed user study showed that ColorPIN is significantly more secure, but entering a four digit PIN took about 12 seconds.

One of the first shoulder surfing resistant authentication schemes for the input of graphical secrets was proposed by Wiedenbeck et al. [278]. The convex hull scheme is based on the recognition of multiple pass-icons. The pass-icons are randomly positioned within a larger set of decoy icons. Instead of directly clicking on the icons, the user is required to click somewhere inside the area which results from mentally connecting the outer borders of the pass icons (as indicated in Figure 2.8 D). As this procedure must be repeated several times and each round comprises a complex search task, users needed over one minute to successfully authenticate. Zhao and Li [294] designed a similar system for the entry of alphanumeric passwords. Khot et al. [147] designed WYSWYE (Where You See is What You Enter) to allow the indirect input of recognition-based secrets. Using WYSEYE, the user identifies her previously memorized pass icons in a larger grid and maps the identified patterns of these pass icons into a smaller grid (see Figure 2.8 E). As the user thereby mentally eliminates the rows and columns which do not comprise pass icons, the process is very hard to follow for observers. Nevertheless, a controlled lab study indicated long authentication times (> 100 seconds) and high mental load. Altiok et al. [6] proposed Graph Neighbors where users are required to memorize a symbol, a color and a direction. To authenticate, the user has to find the secret which matches the previously memorized shape and color and then clicks on a neighbor of this symbol. The respective neighbor is indicated by the memorized direction. The authors tested different versions of the system providing different complexities and reported success rates from 74% to 96%. Input times and empirical shoulder surfing data was not provided. Ho et al. [130] published a recognition-based authentication system which is based on the indirect selection of a specific sequence of images. To authenticate, the user clicks on a target picture which is derived from a starting picture and a cueing picture. In contrast to the starting and the cueing pictures, the targets are generally not part of the user’s pass images. The evaluation of the system indicated authentication times over 50 seconds.

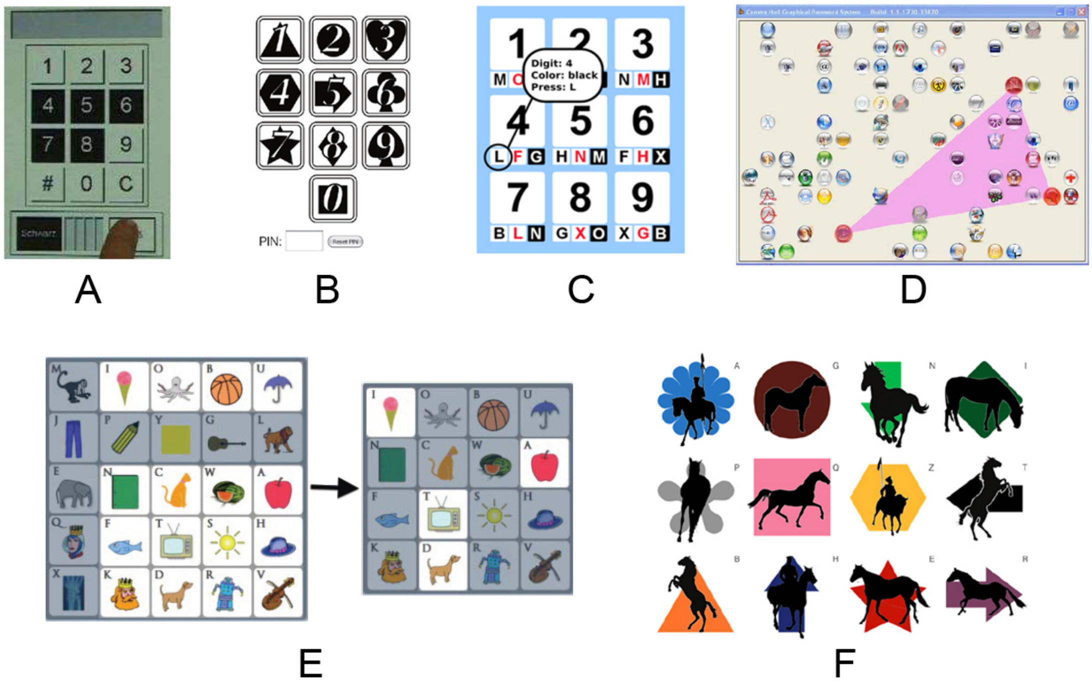


Figure 2.8: Observation resistant concepts based on indirect input and multiplexed secrets. (A) Cognitive Trapdoor Games [207], (B) FakePointer [244], (C) ColorPIN [72], (D) Convex Hull Scheme [278], (E) WYSWYE (Where You See is What You Enter) [147] and (F) Picassopass [258]

In addition to indirect selection, several concepts make use of multiplexed input. In such concepts, the user never selects a single token, but touches a set of objects comprising the secret. Therefore, these concepts have their origin in the cognitive trapdoor games by Roth et al. [207]. The hybrid authentication scheme by Gao et al. [105] which was discussed in Section 2.2.4 is a representative of this concept class. As the user draws a line through both pass-icons and decoy images, the actual secret remains hidden in the set of activated icons. Based on the same idea, Kita et al. [152] designed the Secret Tap Method. The user interface presents 16 icons, the user repeatedly indicates a subset of four icons comprising his secret icon. In addition, a predefined shift value allows the indirect selection of the icon set. Bianchi et al. [26] proposed ShaPIN, where secrets are composed of an arbitrary sequence of numbers, letters, colors and shapes. During authentication, the graphical user interface presents different combinations of these elements. The user selects the button which comprises the intended element. The authors reported authentication times of over 10 seconds. At the same time, the system was significantly more secure than ColorPIN. van Eekelen et al. [258] presented Picassopass which is, like ShaPIN, based on the layered combination of different graphical elements. As depicted in Figure 2.8 F, each token is a combination of a basic shape, a color, a letter and a shape based on a specific theme (e.g., horses). The specific character or the used elements shall support story-based secret selection. The authors claim that multiple observations are required to expose the entered secret. As any selection of such multiplexed tokens sends variable information to the system and the system has to deduce the user's secret, such concepts do usually not support hashed storage which opens other security holes.

Distraction and Visual Overload

Such concepts confuse potential observers by visually overloading the graphical user interface. Systems of this class are usually not resistant against video-based attacks. However, visually overloading the graphical interface can be a feasible strategy to overcharge an attacker's short-term memory and therefore fends off cognitive observers. Tan et al. [246] designed the Spy-resistant keyboard which allows observation resistant password entry on public touchscreens. The concept breaks the entry of a letter into two distinct phases: the mapping phase and the selection phase. All characters are randomly assigned to a position of the virtual keyboard. As illustrated in Figure 2.9 A, the characters are grouped to sets of three items. In addition, a red line is randomly assigned to one of the characters of each set. If the user wants to enter an character, she needs to find the respective set and then shift the red indicator line until it matches the intended character. In the selection phase, the user starts dragging to the respective position of the character set. At the moment, the interaction starts, the mapping disappears. Hence, an attacker would have to memorize the whole mapping before the selection phase starts to reconstruct the user's input. While a user study indicated good protection against cognitive shoulder surfing, the input time for single characters was more than doubled. De Luca et al. [75] proposed the use of fake cursors to protect mouse-based on-screen password entry via virtual keyboards. The concept introduces multiple dummy cursors which mimic the movements of the real user-operated mouse cursor. As a result, the real mouse cursor is hard to identify for observers (see Figure 2.9 B). On the other hand, the user can easily follow the real mouse cursor as it accurately follows the movements of her hand. A similar approach was published by Watanabe et al. [271].

Zakaria et al. [292] evaluated different techniques to protect the Draw-a-Secret (DAS) mechanism from observation attacks: They tested decoy strokes, disappearing strokes and line snaking. Decoy strokes are based on a similar idea like dummy cursors. While the user draws a line, decoy lines appear on screen and mimic the movements of the input. Please note that this strategy is only feasible for indirect input. Line snaking and disappearing strokes are based the idea that drawn lines disappear and therefore, the observer needs to memorize the whole process. While all three techniques had only medium effects in terms of security, it became apparent that disappearing lines worked better than decoy lines. While the imitation strategy of the decoy lines did not work well enough to distract observers, they irritated the eligible user and therefore downgraded the usability. Gugenheimer et al. [112] proposed ColorSnakes, a gesture-based PIN concept which is also based on fake paths (decoy lines). A ColorSnake password consists of a colored starting digit and four consecutive digits. To authenticate, the user searches the first digit in the memorized color and then indirectly selects the remaining digits by indicating the direction using touch gestures. Figure 2.9 C illustrates the visually confusing user interface. ColorSnakes was evaluated in lab and

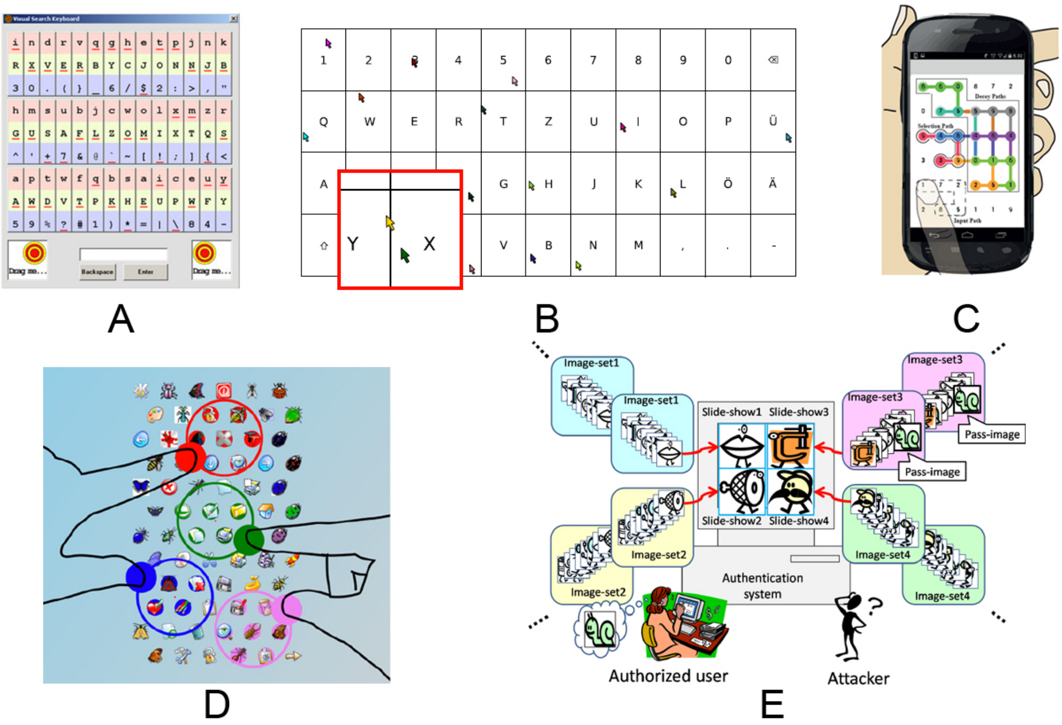


Figure 2.9: Observation resistant concepts based on visual overload and distraction. (A) Spy-resistant keyboard [246], (B) Fake Cursors [75], (C) ColorSnakes [112], (D)Color-Rings [151] and (E) Indirect Image-based Authentication (I-IBA) [285]

field studies and achieved good observation protection, especially when additional countermeasures were taken (e.g., obfuscating the digits after the selection). The authentication times could be reduced to seven seconds after one week of usage.

Kim et al. [151] designed two authentication concepts for tabletops which are based on visual overload. Color-Rings provides four rings to select multiple pass icons. All rings have to be placed at the same time. While only one ring makes the actual input, the other three rings make decoy selections (see Figure 2.9 D). The other concept, called SlotPIN, introduced redundant reels with randomly positioned digits to conceal the PIN entry. Yamamoto et al. [285] proposed a system called I-IBA which simultaneously presents several sideshows. As depicted in Figure 2.9 E, one of the slideshows comprises at least one previously memorized image which needs to be identified by the user. The authors claim that an observer would need multiple video-based attacks to reveal the pass images within a slideshows. Wu et al. [283] combined the idea of the Convex Hull Scheme with elements of dummy cursors. According to their concept, multiple colored balls are randomly moving across the screen. The user has to follow one predefined color, the rest of the balls are decoy elements. Whenever, the predefined ball is moving in the respective area given by the pass-icons, the user presses the space bar. While the concept is robust even against video-based attacks, input times were high (>40 seconds).

2.4 Lessons Learned and Open Questions

The literature review gave important insights into the design, the evaluation and the use of authentication mechanisms. At the same time, it revealed important unsolved problems. In this Section, we summarize the lessons learned and point out open questions.

Usability is an Often Unmet Requirement for Security

Usability is a precondition for feasible authentication concepts. While early work considered usability and security as opponents, the literature review revealed that usability is an indispensable supporter for security. Most security problems discussed in connection with text-based passwords are the result of insufficient usability features. We learned that memorability plays the most important role for text-based passwords: To cope with the large number of accounts, users follow predictable selection strategies and use passwords across multiple services. Graphical and gesture-based authentication mechanisms were shown to reduce the memory burden due to supporting the specific skills of the human brain. At the same time, such concepts are often not feasible due to high input times and are therefore not accepted in the wild.

We conclude that there is usually a trade-off between usability and security, but security cannot exist without usability. This aspect becomes crucial in connection with mobile devices. As users cannot be forced to use security mechanisms, inadequate performance will most likely result in non-usage and therefore completely abolish security. In addition, we assume that the frequent use of mobile devices makes efficiency even more important than memorability.

Theoretical Security is Indispensable but Sometimes Neglected

The discussion revealed that theoretical security is an indispensable requirement for any feasible authentication system. We learned from text-based authentication mechanisms that complex passwords and considerate security behavior become irrelevant as soon as theoretical security fails. This is either the case when clear-text secrets can be accessed by attackers or when a small password space makes exhaustive guessing attacks effective. While disclosed databases provide a great opportunity for password research, they represent a significant security problem for the user. The review of alternative authentication mechanisms showed that the requirement of adequate theoretical security is often neglected. Most recognition-based concepts provide a small theoretical password space and do not support securely encrypted storage. The same is true for gesture-based and drawmetric systems when the matching algorithm does not support password encryption. Section 2.3 revealed that theoretical security is sometimes sacrificed for improved practical security as some concepts achieve observation resistance based on the unencrypted storage of passwords.

We conclude that independent of the context, feasible concepts must not sacrifice theoretical security for improved practical security. Therefore, encrypted storage of secrets must be a precondition for authentication mechanisms on mobile devices.

Practical Security is Vital but Hard to Measure

The literature review indicated an inherent difference between the theoretical protection and the practical security of a system. While theoretical security can be mathematically assessed, practical security is significantly influenced by user behavior and by the context of use and thus requires empirical analysis. After over 30 years of research, the real-world strength of a text-based password is still hard to define. Section 2.2 indicates that the diversity of graphical and gesture-based authentication mechanisms makes the assessment of password strength even more complex. However, as both text-based and alternative secrets were shown to result in a predictable password choice, password popularity was identified as an important security factor. With the increasing number of graphical authentication mechanisms, password capturing (e.g., observation attacks) became another major research topic and various solutions were proposed. Nevertheless, the literature review reveals that despite the large corpus of observation-resistant designs, the practical threat of such attacks and the vulnerability of current concepts are not well understood.

We conclude that it is vital that feasible authentication concepts actually provide the intended security in the wild and thus relevant factors need to be assessed in the actual context of use. For mobile authentication methods, observation-resistance was identified as a major requirement but guessing-resistance is still a crucial feature. Finally, lessons learned from text-based passwords need to be respected when developing novel concepts. While password reuse might be less an issue considering unlock screens, password sharing might become a serious problem [142].

(Mobile) Context Matters but is Not Well Understood

The context is a critical factor for the assessment of the feasibility, security and usability of authentication mechanisms. From text-based passwords, we learned that authentication methods need to fit into the social context of use: Most important, using a specific authentication mechanism should not feel awkward or communicate mistrust. The usage pattern and risk perception are important contextual factors which influence the perceived suitability of a system. Finally, feasible authentication concepts should be tailored to both the target hardware and the target user group to maximize accessibility and acceptability. Mobile devices have special demands which need to be satisfied by feasible authentication systems. Interaction is usually based on rather small touchscreens and frequency of use is high. Authentication may take place in various uncontrolled environments and situations. In addition, the concept must support a diverse set of devices and a heterogeneously skilled target group. The literature review indicates that gestures are well suited for mobile devices: They naturally support touch-based interaction and may even allow eyes-free interaction. At the same time, gestures were also shown to be vulnerable. The literature review revealed that proposed solutions sometimes neglect the special demands of the mobile context. For example, security-optimized designs were often based on additional hardware or resulted in high authentication times. We argue that both aspects do not satisfy the demands of mobile devices: High usage frequency requires efficient authentication tasks and additional hardware seems unhandy and may exclude a huge set of devices.

We conclude that feasible authentication mechanisms need to fit into the context of mobile devices. However, this context comprises various dimension which are not yet understood. It is crucial to understand the usage patterns and the risk perception of users to be able to tailor authentication concepts to the context of use and to meet the users' expectations.

II

Empirical Research

Chapter 3

Exploring the Problem Space of Gesture-based Authentication

All truths are easy to understand once they are discovered;
the point is to discover them.

– Galileo Galilei, astronomer (1564 - 1642) –

Following an inductive research approach [168], we start by analyzing the current state of gesture-based authentication on mobile devices. This Chapter defines the problem space of gesture-based authentication and presents four research projects which aim to understand real-world behavior and real-world problems. The outcome of this Chapter will answer most of the questions that were identified in Chapter 2. In addition, the insights gained from field observation are essential to draw the right implications for future designs.

Section 3.1 defines the problem space and gives an overview of the research covered in this Chapter. In Section 3.2, we investigate risk perception in the wild and gain valuable insights into real-life unlock behavior. Section 3.3 focuses on the comparison of PIN and gestures. The analysis of both concepts in a controlled field study shows important differences in user experience and error handling. Section 3.4 systematically investigates the observability of grid-based gestures and presents a prediction model which can be used to assess the specific risk of a given gesture. In Section 3.5, we present a similarity metric for grid-based gestures and apply it to a corpus of user-defined gestures. The results confirm the vulnerability of current concepts and reveal interesting insights into human preferences regarding geometric properties. Finally, Section 3.6 aggregates the results and draws implications for future designs.

3.1 The Problem Space

The literature review in Chapter 2 revealed a list of requirements for feasible authentication mechanisms. This Section summarizes these requirements and defines the problem space of graphical gesture-based authentication on mobile devices.

The first part of this Section describes the problem space and points out the most important issues. While the problem space could theoretically imply an unlimited number of factors, we focus on relevant aspects that can be improved through enhanced interaction concepts and thorough design decisions. For each factor, we define the goal state (the ideal solution) and assess the current state. The ultimate goal is to achieve a matching of the current state and the goal state [8]. Chapter 4 will discuss potential steps to achieve this research goal.

The second part of this Section gives an overview over *four research projects* that explore the defined problem space and gather important insights into the current state of graphical gesture-based authentication on mobile devices. The current state is mainly assessed utilizing Android unlock patterns, a widely accepted graphical gesture-based authentication mechanisms. In 2015, such grid-based gestures were available on over 80%[1] of all smartphones and consequently represent the perfect test bed to explore real-world problems.

3.1.1 Definition

Chapter 2 shows that feasible authentication mechanisms must provide both *usability* and *security* [1,32]. An ideal authentication system would provide perfect security at zero costs. However, this goal has not yet been achieved and current authentication systems usually comprise a tradeoff between usability and security.

This Section defines the problem space of gesture-based authentication mechanisms on mobile devices. The context is important as the relative weights of usability aspects and security features are likely to change depending on the use case. While most aspects are preconditions for all authentication mechanisms (e.g., effectiveness), others are primarily important for gesture-based authentication (e.g., smudge resistance). Even though single factors are discussed in isolation, it is important to keep in mind that most factors interact with each other: For example, memorability problems may negatively influence practical security.

Usability

Usability is an important factor for all security systems [1]. The usability features determine the direct costs that a user has to pay when using the system. As a consequence, authentication systems which offer insufficient usability will most probably be bypassed. In the context of mobile devices, (perceived) efficiency and effectiveness are exceptionally important for mainly two reasons: First of all, authentication takes place many times a day and therefore,

[1] `http://www.idc.com/prodserv/smartphone-os-market-share.jsp` – last accessed: 2015/11/02.

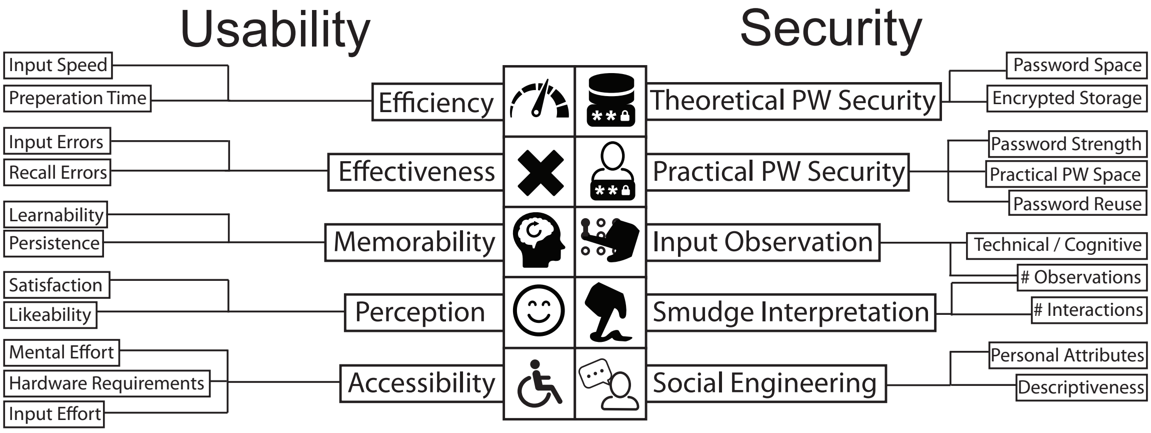


Figure 3.1: The Figure illustrates the most important aspects of the problem space of graphical gesture-based authentication mechanisms on mobile devices.

increased effort quickly adds up to unacceptable costs. Secondly, in contrast to other environments (e.g., ATM), security is optional and therefore authentication systems demanding unacceptable effort will most probably not be used.

Efficiency describes the additional effort a person faces by using an authentication system. While this may imply both *mental effort* and *physical effort*, efficiency is usually assessed by the time it takes to successfully authenticate (e.g., [276]). As authentication time may exceed input time due to *preparation effort* or *clean up phases* (e.g., [112]), it is important to investigate all possible steps of the authentication process.

The ideal authentication system (*goal state*) allows authentication without additional effort. Behavioral authentication mechanisms can theoretically fulfill this requirement as the user is authenticated based on tasks which would be performed anyway [137]. However, such solutions do not support ad-hoc authentication. In the context of mobile devices, efficiency is exceptionally important as authentication usually takes place many times a day. Section 3.2 will show that PIN and Android unlock patterns are satisfactorily efficient for most users while some users still refrain from using unlock screens, claiming that current solutions are too slow. Section 3.3 presents a comparison of unlock gestures and PINs and indicates that *perceived efficiency* is even more important than measured efficiency.

Effectiveness from the usability point of view describes if a user is able to successfully authenticate. Usually, the *error rate* is analyzed to assess effectiveness (e.g., [100]). Errors can be examined in various dimensions: A user may fail in a single authentication attempt or fail completely meaning that the device gets blocked and a fallback authentication system is required [114]. Section 3.3 will show that the number of corrected attempts should be assessed as well, whenever *error correction* methods are provided. Finally, it is important to distinguish the source of errors: While *inaccurate input* indicates interaction problems, *incorrect recall* indicates memorability issues.

The ideal authentication system (*goal state*) eliminates all possible sources of error and guarantees successful authentication. Similarly to efficiency, high success rates are prerequisites for acceptable authentication systems. This is particularly true in the context of mobile devices as authentication is performed in various situations (e.g., walking, talking) and often represents a secondary task. Just like efficiency, the in the wild effectiveness of *current* authentication methods has not yet been evaluated systematically. It is investigated in Section 3.3.

Memorability describes how well secrets (e.g., gestures) are remembered and recalled. Even though memorability is categorized as a usability feature, it directly influences the security of a system [286]. Since long-term evaluations are costly and time-consuming, evaluations are often limited to short-term (minutes) or mid-term (days) recall tasks (e.g., [55]). The memorability of a secret is mainly influenced by its *meaningfulness* and the *frequency of use* [40]. That is, a frequently used secret is easier to recall than a rarely used one and a self-selected and meaningful secret can be easier recalled than an abstract and system-assigned one. In addition, *internal processing* (visual, motor, linguistic) and the *retrieval strategy* (recognition, recall) play an important role [32]. In addition to the persistence of a learned secret, the learnability itself is an important aspect: It describes the required effort to learn a new secret or a new authentication concept.

The ideal authentication system (*goal state*) empowers users to create and to immediately memorize an infinite number of different secrets. In reality, there is an upper bound for memorized tokens. Even if memorability can be assumed to be a minor problem in the context of mobile devices as unlocks are usually performed many times a day, the memorability of *current* unlock mechanisms has not yet been evaluated systematically. Section 3.3 presents first insights and indicates good memorability for both PIN and unlock gestures.

Perception is a critical factor for the usability assessment of authentication systems. It is usually measured through user feedback collected via questionnaires and interviews (e.g., [73]). User perception can be distinguished into *satisfaction* and *likeability*. The first aspect describes how users perceive efficiency, effectiveness, memorability and security of a given system. The second aspect deals with user experience and describes if people enjoy the system. Both aspects are strongly influenced by the context of use and by social factors [1]. For example, the perceived risk may determine how much people are willing to pay (perceived effort) for additional security [23].

The ideal authentication system is both satisfying and joyful to use (*goal state*). In addition, this should only be the case if the system is used in its most secure configuration. To design feasible authentication mechanisms which people will use, it is important to understand how risks are perceived and how unlock mechanisms are used in the wild. Section 3.2 provides insights into real-world behavior while Section 3.3 sheds light on the perception of PIN and unlock gestures.

Accessibility is important to support a maximum range of users in a maximum extent of situations. The most important factors are *mental effort*, *physical effort* and *hardware requirements*. Increased mental effort or high physical demand might exclude a huge portion of users (e.g., elderly people). The same is true for authentication mechanisms with special hardware requirements. Even though "accessibility" is often discussed in connection with disabled persons, we define accessibility as an important aspect for all users.

An ideal authentication mechanism (*goal state*) supports all users in all situations. Considering mobile devices, the accessibility of an authentication mechanism is often determined through its touchscreen feasibility. In addition, eyes-free interaction and one-handed input can increase the accessibility in certain situations. *Currently* deployed knowledge-based authentication mechanisms are accessible for most users. However, improved security features often reduce accessibility due to increased effort or due to the requirement of specific hardware. When designing novel authentication systems, it is important to keep in mind that feasible concepts must improve security without giving up accessibility. Section 3.2 gives insights into some of the barriers people face with current authentication systems. In addition, accessibility is a prerequisite for all systems discussed in Chapter 4.

Security

Security is the essential factor for authentication systems and the only reason why an authentication system is used. However, the actual security level of a system depends on the used threat model and remains remarkably hard to define. In contrast to usability costs, security benefits are hardly directly perceived. As a consequence, users tend to sacrifice security for usability. Therefore, it is indispensable to assess the practical security which is influenced by the user's real-life behavior. In the context of mobile devices, practical security is mainly determined by two factors: (1) by the practical password space that describes the vulnerability to guessing attacks and (2) by the risk of exposing the used secret to potential attackers.

Theoretical Password Security is mainly defined by the maximum number of available secrets (theoretical password space). A small theoretical password space allows exhaustive *brute force* attacks. Furthermore, a feasible concept needs to support securely encrypted storage of secrets to prevent eavesdropping. Both aspects are determined by the design of an authentication mechanism (e.g., composition rules, available elements, matching algorithms) and not influenced by user behavior.

The ideal authentication system (*goal state*) provides an unlimited pool of secrets. However, recent tests[2] indicate that adequate protection is already achieved by providing 10^{18} different secrets. To provide such a large theoretical password space, PIN users would need to memorize 18 digits. In real life, most PINs are based on four digits which corresponds to a significantly smaller space of 10,000 tokens. The upper bound of Android unlock gestures is limited by design. Overall, it provides only $389,112$ different gestures [18] and can be exhaustively searched within seconds [200]. We therefore conclude that *current* methods

[2] `http://cgi.distributed.net/speed/` – last accessed: 2015/10/28

do not provide adequate brute force protection by themselves. However, mobile devices usually prevent brute force attacks by disallowing larger numbers of successive authentication attempts. Nevertheless, a minimum number of $n > 10^4$ secrets will be defined as the precondition for the design of novel concepts.

Practical Password Security is determined by the *strength* and the *diversity* of the passwords which are actually in use. The practical password space is usually significantly smaller than the theoretical password space as users often opt for secrets which are both easy to remember and easy to use [38]. The practical password space is the main security measure to assess the vulnerability of a system according to educated guessing attacks (e.g., dictionary attacks). The strength of an individual secret is usually hard to define but it is widely accepted that the popularity of a chosen password plays an important role [167].

The ideal authentication system (*goal state*) provides an infinite pool of secrets that are all equally easy to remember and equally easy to use. As a consequence, users select any possible secret with the same chance and the practical password space matches its theoretical size. However, *current* knowledge-based authentication systems comprise a tradeoff between password composition and password usability. First results indicate that the selection of Android unlock patterns is very predictable [254]. Section 3.5 will provide a systematic evaluation of pattern similarity and shed light on user preferences and gesture strength.

Observation Resistance describes the resistance against both *internal capturing* (e.g., touch loggers) and *external observations* (e.g., shoulder surfing). The threat of internal capturing is usually independent from the used authentication mechanism and can be prevented by keeping the operating system up to date and by performing regular malware scans [127]. In contrast, external observability is determined by the design of an authentication concept (e.g., [207]). The level of resistance is significantly influenced by the interaction style and by the interface design. Observation attacks are categorized into *cognitive* and *technical* (e.g., camera-based) approaches[3]. Furthermore, the *number of observations* plays an important role as some concepts are resistant to single observations but vulnerable to multiple attacks (e.g., [244]).

The ideal authentication system (*goal state*) makes it impossible to derive the entered secret from observing the input. Observation resistance is an important security feature for mobile authentication as interaction often takes place in public and uncontrolled environments. *Current* knowledge-based authentication systems seem to be prone even to single cognitive observation attacks. However, the real-world risk level and the impact of composition factors have not yet been evaluated. Section 3.2 will investigate the perceived risk of observation attacks, Section 3.4 will analyze the vulnerability of Android unlock patterns and present a prediction model to assess the security of a given gesture.

[3] "observation" can also describe other modalities than sight (e.g., with a microphone)

Smudge Attack Resistance is a security factor which becomes crucial with touch-based interaction. In contrast to direct observation attacks, smudge attacks are not bound to the act of authentication and can be performed any time, especially in the absence of the device owner [18]. As smudge traces are constantly overwritten and modified with every touch, the threat of smudge attacks drops with the *number of touches* that are performed after a secret was entered. Similar to observation resistance, the *number of attacks* plays an important role as different parts of a secret might be interpretable at different stages.

The ideal authentication system (*goal state*) leaves smudge traces in a way that makes interpretation impossible. In the context of mobile devices, smudge attacks are a critical threat since interaction usually takes place via touch input. *Currently* used gesture-based authentication methods were already shown to be very prone to such attacks, they even allow interpreting the temporal order of the input [18]. So far, smudge attacks have only been evaluated in lab experiments simulating worst case scenarios[4]. The real-world threat of smudge attacks in the users' daily life is still unknown. Chapter 4 will present several interaction concepts which leave smudge traces that are harder to interpret.

Divulgation Resistance describes the level of protection concerning the manipulation of users (social engineering). Even if it is an important security factor of authentication systems, it is not limited to the context of passwords and may address the disclosure of any private information. Attacks can be categorized into *technical approaches* (e.g., phishing) and approaches that are based on *individual interaction* [1]. The vulnerability to individual attacks is influenced by the characteristics of the used passwords as secrets which are easy to describe or based on personal attributes comprise a higher risk to be divulged.

The ideal authentication system (*goal state*) prevents any kind of unintended divulgation. Mobile unlock mechanisms are usually protected from phishing attacks as the authentication process is physically bound to the mobile device. Phishing attacks may, however, become a problem when secrets are reused for remote services. Since gestures are not based on personal attributes or natural language, gesture-based authentication mechanisms usually comprise lower risk for unintended divulgation than text-based passwords. However, the vulnerability of Android unlock patterns to social engineering attacks (*current state*) has not yet been evaluated.

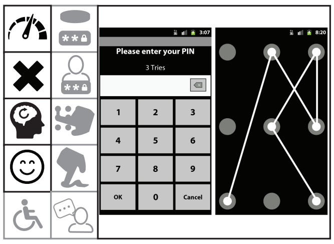
[4] The authentication is performed on a clean surface and no further interaction takes place.

3.1.2 Overview

The remainder of this Section presents four research projects which were conducted to explore the current state of gesture-based authentication on mobile devices. The first and the second project mainly deal with usability issues and provide important insights into real-world efficiency, effectiveness, memorability and perception. The third and fourth projects focus on security and investigate the practical password space as well as the observation resistance of current gesture-based authentication systems.

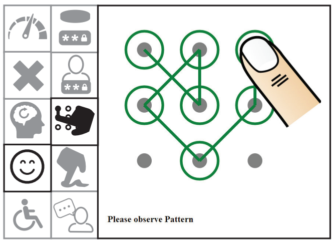
Unlocking Behavior and Risk Perception in the Wild

Section 3.2 sheds light on risk perception and unlocking behavior in the wild. The evaluation is based on an online survey (n = 260) and on a field study (n = 52). The results reveal important insights into the real-world efficiency of current solutions. We learn how users perceive often discussed threats and for what reasons people decide to or not to use security mechanisms.



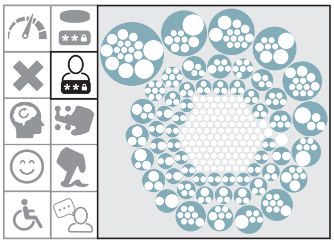
Usability of Grid-based Gestures in the Wild

Section 3.3 gives empirical evidence for the real-world usability of gesture-based authentication systems. The evaluation is based on a controlled field study (n = 60) which compared gestures to PIN and collected both qualitative and quantitative data. The results reveal important insights into the relationship of perceived usability and measured performance data.



Observation Resistance of Grid-based Gestures

Section 3.4 quantifies the observation resistance of current gesture-based passwords, that is of Android unlock patterns. The evaluation is based on 5960 observation attacks which were simulated in an online study (n = 298). The results reveal that Android unlock patterns are prone to such attacks but at the same time the composition strategy has a significant impact on observation resistance.



Practical Password Space of Grid-based Gestures

Section 3.5 quantifies the practical password space of current gesture-based passwords, that is of Android unlock patterns. A novel similarity metric is introduced and applied to a dataset of 506 user-defined unlock patterns. The results reveal that most users prefer very similar shapes and indicate that currently used secrets are vulnerable to dictionary attacks.

3.2 Unlocking Behavior and Risk Perception in the Wild

Chapter 2 illustrates that a lot of research is being conducted to improve the status quo of (mobile) authentication. At the same time, little is known about users' attitudes towards currently deployed solutions (e.g., PIN). As novel concepts are usually evaluated in controlled lab experiments where current methods only serve as a baseline condition, we still lack important insights on users' unlocking behavior and risk perception in the wild. However, a profound understanding of these factors is a precondition to design feasible authentication mechanisms which fit in the context of use and meet the expectations of the user. This Section will define the basic requirements for feasible mobile authentication concepts by shedding light on the following *research questions*.

RQ1 Why do or do not users lock their phone?

RQ2 How often and in which situations do people use secure lock screens?

RQ3 How often and in which situations do people access sensitive data?

RQ4 How do people perceive the risks of observation or unwanted access?

We based the evaluation on an *online survey* (n = 260) and on a *field study* (n = 52). While the online study gave a broad overview of general locking motivations and protection strategies, the four-week field study allowed analyzing risk perception and user behavior while actually interacting with unlock mechanisms. Overall, we captured one month of locking activities and sampled 6582 in-situ experiences. This Section contributes to the understanding of risk *perception*, sheds light on *satisfaction* and delivers insights into real-world *efficiency* features of current solutions.

We will learn that the decision for not using a secure lock screen is often based on reasonable justifications and as a consequence unlock mechanisms must provide superior usability to get adopted (*RQ1*). This is especially true since the specific pattern of use is described by frequent unlocks combined with brief periods of use (*RQ2*). Even with PIN and pattern, we found that users were often dissatisfied and desired the presence of Slide-to-Unlock. In addition, the results reveal that sensitive data is seldom accessed (*RQ3*) which questions the necessity of a good portion of unlock operations. Finally, we show that often discussed risks like observation attacks or unwanted access were rarely perceived (*RQ4*).

This Section is partly based on Harbach, M., von Zezschwitz, E., Fichtner, A., De Luca, A., & Smith, M. (2014, July). It's a hard lock life: A field study of smartphone (un)locking behavior and risk perception. In Proceedings of SOUPS'14 [118]. Please refer to the beginning of this thesis for a detailed statement of collaboration.

3.2.1 Research Context and Motivation

It has already been shown that individual goals and perceptions are significant factors for the adoption and the use of security mechanisms [1]. According to Beautement et al. [23], users have a limited "compliance budget" and permanently evaluate the cost-benefit tradeoff of using a security system. Security systems which are rated to comprise more costs than benefits are usually not accepted. The perceived costs which mostly imply additional effort (e.g., physical load, cognitive load) are directly perceived. The benefit is, however, difficult to quantify and depends on previous experiences and risk perception [128].

Risk perception is generally hard to measure and influenced by various factors [231]. Becher et al. [24] claim that mobile devices comprise specific risks which differ from traditional devices. Felt et al. [92] conducted a large-scale survey on smartphone users' concerns and defined 99 potential risks (e.g., phone damage, location sharing). Chin et al. [58] revealed that privacy risks are perceived more critical when using mobile devices. Theoharidou et al. [250] suggest various quantitative (e.g., loss, theft) and qualitative (e.g., integrity, availability) factors to assess the practical risk of smartphones. Muslukhov et al. [185] investigate user practices in protecting mobile device data and reported that existing solutions require too much effort. Van Bruggen et al. [257] found out that over 30% of the users refrain from using authentication on mobile devices. Finally, Mylonas et al. [186] present a taxonomy for sensitive data found on mobile devices and assess the actual risk of a threat based on the combination of likelihood and estimated consequences.

Despite such risk assessment, usage patterns of mobile device users have already been quantitatively analyzed. Andrews at al. [9] report that the average user checks the smartphone 85 times a day and that even though most sessions are shorter than 30 seconds, the overall usage time adds up to five hours each day. Furthermore, mobile device use has been reported in non-academic studies. A recent Gallup panel survey[5] found that 52% of the users in the US check their smartphone at least once an hour. The Nielsen Company[6] reports a monthly use of 37 hours and 28 minutes. The app company Locket[7] found that users unlock their devices 110 times a day. Finally, specific applications have been released which allow users to monitor their own usage behavior (e.g., checky[8]).

In addition to the prior work mentioned above, relevant work has been published after the herein reported results had been released. Volkamer et al. [261] performed semi-structured interviews to assess the justifications for the non-use of security measures and confirmed most of the findings presented here. De Luca et al. [70] analyzed why people use or not use

[5] `http://www.gallup.com/poll/184046/smartphone-owners-check-phone-least-hourly.aspx` – last accessed: 2015/11/11.

[6] `http://www.nielsen.com/us/en/insights/news/2015/so-many-apps-so-much-more-time-for-entertainment.html` – last accessed: 2015/11/11.

[7] `http://www.npr.org/blogs/alltechconsidered/2013/10/09/230867952/new-numbers-back-up-our-obsession-with-phones` – accessed 2015/11/11.

[8] `http://www.checkyapp.com` – last accessed: 2015/11/11.

biometric unlock screens. The results motivated Micallef et al. [181] to evaluate the usability of context-sensitive concepts. Finally, Egelman et al. [85] analyzed the reasons for locking or not locking smartphones and investigated more qualitatively how lock screens are used.

3.2.2 Online Survey

The results of the online survey particularly shed light on users' locking motivations, risk perceptions and performed countermeasures.

Method

The survey was distributed using Amazon's Mechanical Turk (MTurk). Participants were required to use a smartphone regularly for at least three months. MTurk has been shown to be suited for usable security evaluations [145] if specific precautions are taken [79]. To control the inclusion criteria, the ownership of a smartphone had to be proven by opening a provided link using a mobile device. The link checked if the HTTP user agent string matched a known mobile browser. In addition, we included several attention check questions to validate participants' carefulness. Participants who passed both the device check and the attention check were paid $0.70 of compensation.

The survey consisted of three main parts: We investigated risk attitudes, extra measures and critical incidents. In addition, we collected demographic information and inquired IT experience. Answers concerning extra measures, locking motivations and critical incidents were collected using open ended questions and were inductively coded by two researchers. The preliminary code plans were then discussed and merged. Finally, both researchers coded the responses, possibly assigning multiple codes to each statement. Conflicting decisions were discussed and resolved before a third coder independently coded all responses using the final code plan.

Participants

From originally 320 participants, we had to remove 60 subjects due to incorrect completion codes or due to wrong answers to two or more attention checks. On average, participants were 33 (SD = 10; 18-67) years old, 54.6% were male. Most (74%) participants were employed, 9% were students and 17% indicated unemployment or “other”. Almost a quarter of the participants indicated an IT background and 40% reported a very good understanding of computers. On average, participants had used smartphones for 35 (SD = 21; 3-144) months. All subjects reported daily use and the majority (79%) checked their smartphones at least once per hour. 49% used iOS devices, 48% used Android and 3% used other operating systems. 51% reported to have experienced smartphone related privacy or security incidents before. Most (30%) of them reported device loss (or theft), 29% reported a broken phone or data loss and 12% reported unwanted access. Overall, 43% of participants indicated to use secure lock screens, including PIN (78%), unlock patterns (20%) and passwords (2%). Furthermore, 41% used Slide-to-Unlock and 16% had no lock screen activated.

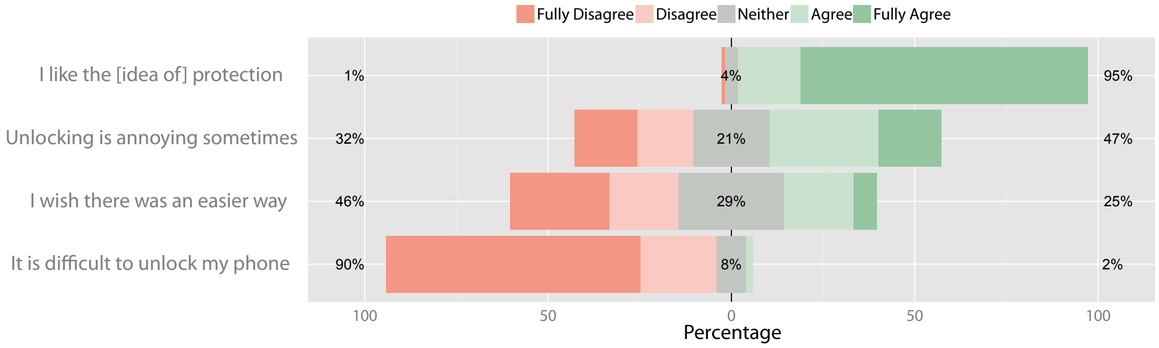


Figure 3.2: Users of secure lock screens indicated their attitudes based on 5-point Likert scales.

Results

Satisfaction with current lock mechanisms was evaluated using 5-point Likert scales. The results are depicted in Figure 3.2. We focused on the 111 users that reported to use secure locking mechanisms. 95% of the participants stated that they liked the idea that the unlock screen protects from unwanted access. While only 2% agreed that unlocking the phone is difficult, 47% still somewhat or fully agreed that it can be annoying sometimes. In addition, 25% somewhat or fully agreed that easier authentication mechanisms were desired.

Motivation for using or not using a secure lock mechanism was evaluated based on open ended questions. The use of lock mechanisms was most often motivated by general protection goals (79%) like "*access control*" (29%) or "*security*" (23%). In addition, 68% wanted to protect certain information (e.g., photos), 56% mentioned specific scenarios (e.g., loss) and 50% referred to attackers (e.g., room-mates, children). In particular cases, the use of secure lock screens was the consequence of external factors (e.g., required by the company).

NOT using a secure lock mechanism was mostly (79%) justified by the absence of threats. 17% indicated "*no need for security*", 15% argued "*they had nothing to hide*" and 11% said they "*did not store any sensitive data*". Interestingly, 19% of the participants explained the absence of threat with physical protection: "*My phone is always right beside me or in my pocket [..]*". Beyond that, inconvenience was an important factor for 57% of the users: 15% specifically stated that it took "*too much time*", 9% said they would use the "*device too frequently*". Finally, some participants reported specific barriers like device sharing, emergency situations or reduced chances of being contacted in case of hardware loss.

Risk perception concerning *mobile device threats* was assessed using 5-point numeric scales. Most (65%) participants were not or mostly not concerned about someone observing their unlock. In addition, we provided a list of six common incidents and asked participants to indicate the worst scenario. Overall, 53% identified "*losing the phone itself, because I would have to buy a new one*" as the worst scenario. This indicates that most users value hardware costs higher than privacy and security risks. Data loss was selected by 20%, 12%

chose account abuse as a consequence of a lost phone and 9% were most afraid of data abuse on a lost phone. Finally, 4% indicated app abuse on an unattended phone as most critical and 1% was most afraid of data abuse on an unattended phone.

In addition, we assessed the risk perception concerning *specific attackers*. First, participants rated the probability of known malicious, known curious, unknown malicious and unknown curious attackers on a 5-point scale. Known curious and unknown malicious attackers were considered more likely than the other two types. In a second step, all participants who had rated known attackers as neutral to likely were asked to assign one of the following aspects to eight presented types of known persons: ‘‘*potentially curious*”, “*potentially malicious*” and “*I did not consider this group of people*”. The results reveal that closely related persons are mostly considered as curious. Close friends were most often (73%) reported. Furthermore, 54% of the participants selected acquaintances, 53% parents, 52% children and 46% friends of friends. In contrast, more remotely known people were considered as potentially malicious attackers. In this category, “*other known people*” (67%), co-workers (29%), acquaintances (25%) and friends of friends (23%) were most often chosen.

Extra measures besides using a secure lock screen were analyzed based on free text and predefined answers. When we asked participants if they sometimes take one or more specific countermeasures, 84% selected that they “*conceal [the] smartphone in [their] clothes or in a bag*”, 51% indicated to leave the device at “*a safe place before going somewhere*” and 34% “*enable a lock screen for this situation or choose a harder PIN/password/pattern*”.

Furthermore, we asked for up to three other situations in which additional measures are taken. “*Paying extra attention*” was mentioned most often (45 instances), but also technical measures (e.g., turning off the device) and physical extra measures (e.g., privacy foil) were indicated. Concerning risky environments, participants mentioned public and semi-public spaces and situations like going “*out*” (59), during parties (39) or at work (52). In addition, dangerous neighborhoods were frequently mentioned (24). Only 16 participants reported private spaces (e.g., home). Despite location-dependent events, participants reported generally unattended situations (71) and phases of less caution (102). Finally, person-related situations like unfamiliar or untrusted persons (20), kids (9) or (ex-)partners were mentioned.

In addition, we asked for extra measures which were taken against shoulder surfing. We provided five answers and allowed multiple selections. Most (28%) participants indicated that they tilt their screen away while entering their unlock code, 16% claimed to wait a moment before entering the secret, 11% turn around and 9% cover the display. Only 7% acknowledged to have changed their unlock code after a potential shoulder surfing attack.

Critical incidents related to unwanted access were reported by 31 participants. Due to the small sample size, the answers were not coded. Participants mentioned snooping (ex)partners as well as children or siblings fooling around with the devices. In addition, friends playing pranks and thievery were mentioned. Reported harm included invasion of privacy, conflicts with other persons and account abuse.

3.2.3 Field Study

The online survey provided valuable insights into users' reasoning concerning the use or non-use of lock screens. Nevertheless, the data was based on retrospective self-reports and therefore influenced by participants' memory. To collect unfiltered real-world data of unlock behavior and risk perceptions, we additionally performed a four-week field study (n = 52). The study design was informed by the results of the online survey and allowed to answer the main research questions while ensuring ecological validity.

Method

The main goal of the field study was to evaluate real world behavior. Therefore, we needed a user study design which allowed unobtrusive data collection and would most probably not influence user behavior. For this purpose, we developed an application which collected quantitative information about the unlock behavior. In addition, we followed an experience sampling approach similar to Cherubini and Oliver [51] and displayed mini-questionnaires after certain unlocks to collect qualitative data. In a first meeting, we helped to install the application and explained the study goal as well as the procedure. After the study was finished, participants were invited for debriefing. We performed a short interview to assess if the user study had an impact on the observed behavior and to gain insights into potential problems. Finally, we helped uninstalling the study app. The following provides detailed information about the most important aspects.

Unlock behavior was analyzed based on four events: *activation*, *unlock*, *lock* and *deactivation*. Whenever a new state was entered, a timestamp was logged. The state model comprised four states:

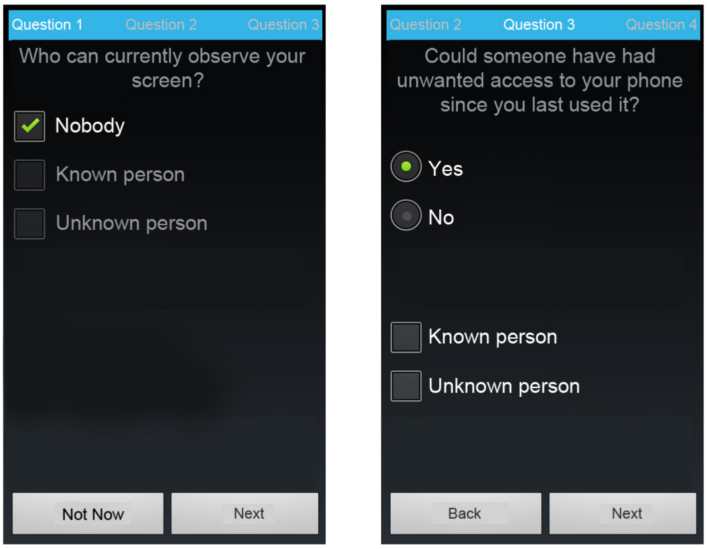
Off Locked The display is off and the device is locked.

On Locked The display is on and the device is locked.

On Unlocked The display is off and the device is unlocked.

Off Unlocked The display is off and the device is unlocked.

Transitions are possible between all four states. Transitions from *Unlocked* to *Locked* are mainly depending on the device configurations: Common configurations are "lock immediately" or "lock after x minutes". The same is true for screen-off events (*Off*). Screen-on events (*On*) are usually triggered by user interaction (e.g., button presses) or might be triggered by the operating systems (e.g., incoming call). The transitions from *Locked* to *Unlocked* require interaction with the unlock screen (e.g., Slide-to-Unlock, PIN, Pattern). Logging the transitions between the four states allowed a detailed analysis of the users' unlock behavior. However, it is important to note that the unlock times computed with this model represent a worst-case estimate as they may include time-consuming interaction performed on a locked screen (e.g., viewing notifications).



Who can currently observe your screen?

☑ Nobody

□ Known person

□ Unknown person

Could someone have had unwanted access to your phone since you last used it?

● Yes

○ No

□ Known person

□ Unknown person

Figure 3.3: Two examples of the graphical user interface used in the field study. While one questionnaire (*unlock risk survey)* focused on the unlock process (left), the other questionnaire (*data risk survey*) investigated the time frame between the current unlock and the previous one (right). The questions were originally posed in German.

Mini-Questionnaires were designed and optimized for quick and easy interaction. Figure 3.3 illustrates two examples of the graphical interface. The questionnaires were both based on seven multiple choice questions and randomly displayed after device unlocks. One questionnaire, called *unlock survey*, focused on the actual unlocking procedure. It investigated the perceived shoulder surfing risk, additional measures and the satisfaction with the performed unlock. In addition, we asked participants to rate the sensitivity of the accessed data. The second questionnaire, called *data survey*, focused on the time span between two unlocks. Participants were asked to indicate if additional measures were taken since the last unlock and reported the risk that someone else could have had unwanted access. Finally, both questionnaires ended with an assessment of the current environment based on three categories: private, semi-public and public.

Sample Rate plays a significant role for the outcome of the study as it comprises the main trade-off between the additional effort and the amount of collected data. A too high sample rate is likely to cause negative effects in terms of changing the observed user behavior and would therefore decrease ecological validity. A too low sample rate would exclude a huge number of interesting day-to-day situations and lead to an invalid data collection. We decided to randomly select unlock events after which one of the two questionnaires was presented. In addition, we added a mandatory break of one hour after a questionnaire was filled in. Based on pre-study results, we started with a probability of 20% for all users. After one week, we adjusted the sample rate to the actual user behavior. We therefore categorized the participants into three groups and chose a sample rate which resulted in five to six presented questionnaires per day. Heavy users with at least nine unlocks per hour were sampled with a

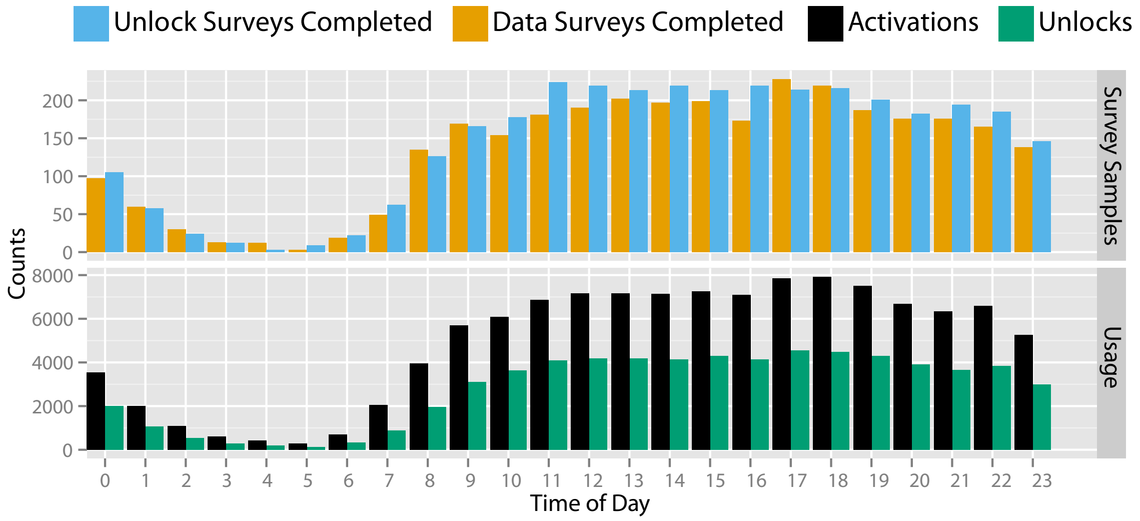


Figure 3.4: This Figure maps the collected data to the time of the day. Overall, we logged 116601 activations, 66874 unlocks, 3410 unlock surveys and 3172 data surveys.

chance of 10%, users with four to eight unlocks per hour received an update to 15%, all other participants remained at 20%. In addition, users were allowed to dismiss questionnaires by pressing a “Not Now” button (as indicated in Figure 3.3, left.).

Participants

The user study was performed at two locations in Germany. In Munich, we recruited 27 participants through mailing lists and social media. In Hanover, 30 participants were recruited using a specific study participation mailing list. All participants owned a smartphone with Android 2.3 or higher and had used it for at least three month. For compensation, we offered a 10 Euro base-salary plus 14 Euro-cent per completed mini-questionnaire.

Five participants had to be removed due to logging errors (n = 4) and due to missing out the debriefing (n = 1). The average age of the remaining 52 participants was 24 (SD = 3; 19-32) years. 29 were male and 23 were female. The sample was skewed towards high education as 47 participants indicated to be (under)grad students and the remaining five subjects were PhD students or academic employees, 48% indicated an IT background. All participants were experienced smartphone users with an average prior use of 34 months (SD = 20; 5-120). The sample comprised the most important lock mechanisms: 13 PIN users, 22 pattern users and 17 Slide-to-Unlock users. PIN users indicated an average PIN-length of 4.5 digits (SD = 0.8; 4-6), pattern users based their secrets on 5.2 cells (SD = 1.3; 4-8) on average.

Results

Each data set was truncated to the first 27 complete days ranging from midnight to midnight. Since unlock behavior naturally differs between subjects and the sample rate was influenced by the number of unlocks, each participant contributed a different number of data points. To counteract over-representation of heavy users, data sets were aggregated on user level and analysis was based on the average across these aggregates whenever appropriate.

In 27 days, we logged an average of 2242.3 activations per participant (SD = 1160.2, 651-5419). 1286.0 (57.4%) of these screen on events were followed by a device unlock (SD = 711.8, 215-3545). In addition, each user contributed an average of 65.6 unlock risk surveys (SD = 3.0; 15-110) and an average of 61.0 data risk questionnaires (SD = 2.7; 15-105). Figure 3.4 illustrates the collected data according to the 24 hours of the day. The mapping indicates that indeed all times of the day were sampled and that the number of samples seems to match common daily routines (e.g., less activity at night time). In the following, we present the results considering effort, context, sensitive data and risk perception.

Effort and Session Length

On average, participants checked their mobile devices (*screen on*) 83.3 times a day (SD = 43.0, 24.2-201.1). 57.4% (Mn = 47.8, SD = 26.4, 8.0-131.5) of these device checks included a device unlock. Figure 3.4 shows that users followed daily routines and were most active between eight in the morning and midnight which at the same time indicates an average sleep of eight hours. Therefore, we assumed that the average user is awake 16 hours per day. This translates to a device activation every twelve minutes (Mn = 5.2/h, SD = 2.7) and an unlock every 20 minutes (Mn = 3.0/h, SD = 1.7). Figure 3.6 (left) indicates a bimodal distribution of unlock events. We find that half of the users unlock their device up to three times per hour, while the rest unlocked the device more than three times including heavy users which unlocked their device at least every ten minutes. Within the first meeting, we asked participants to estimate their average number of unlocks. Figure 3.5 compares these pre-study guesses with the actual numbers derived from the log data. The analysis shows that most users underestimated the daily number of device unlocks.

In addition to unlock frequency, unlock time plays an important role when assessing unlock effort. On average, participants needed 2.67 seconds (SD = 8.46s) using Slide-to-Unlock, 3.0 seconds (SD = 13.3) using a unlock patterns, and 4.7 seconds (SD = 20.72) using a numeric PIN. For exploratory investigation, we grouped participants post-hoc into regular users (unlocks/h$\leq$3) and heavy users (unlocks/h>3). A user-type $*$ lock-screen between-subjects ANOVA based on the average unlock times per user found a significant main effect for lock-screen ($F(2,46) = 11.37$, $p < .001$) and a significant main effect for user-type ($F(1,46) = 6.39$, $p = .002$). A Bonferroni-corrected pairwise t-test revealed that heavy users spent significantly less time on unlocking (Mn = 2.9 sec, SD = 1.2) than regular users (Mn = 3.8 sec, SD = 1.6), $p < .05$. In addition, PIN users (Mn = 4.9 sec, SD = 1.9) performed significantly slower than both pattern users (Mn = 3.2 sec, SD = 0.9, $p < .001$) and Slide-to-Unlock users (Mn = 2.6 sec, SD = 1.0, $p < .001$). However, no significant differences

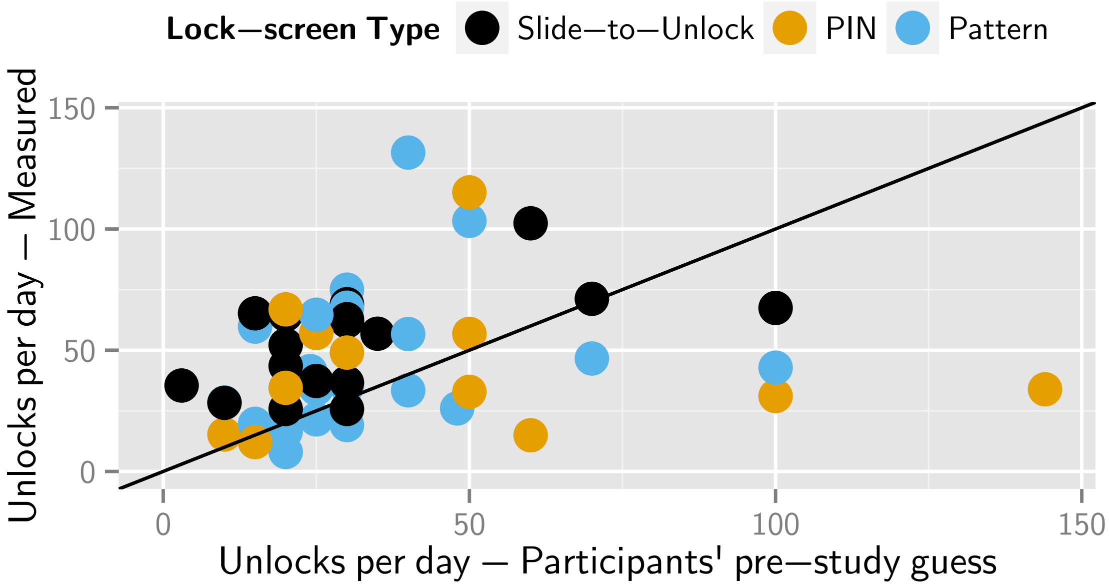


Figure 3.5: The comparison of guessed and actual activation numbers indicates that most users underestimate unlock frequency.

were found between pattern and Slide-to-Unlock ($p > .05$). Over the course of 27 days, participants spent an average of 1.17 hours (SD=.87, 0.2-5.1) unlocking their devices.

We assume that the perceived unlock effort is related to session length as the relative overhead caused by the lock screen grows with decreasing session times. Analyzing the time between all *Screen ON* and *Screen OFF* events indicates an average session length of 71.1 seconds (SD = 245.4). The average session length on locked devices was 12.8 (SD = 315.2) seconds, on unlocked devices sessions lasted 104.5 (SD = 186.2) seconds on average. Figure 3.6 (right) illustrated the session lengths for both states. The black bars which illustrate the session length on locked devices show a bimodal distribution. We assume that the first peak at approximately one second results from common short tasks (e.g., checking the time) while the second peak may be the result of longer interactions (e.g., reading notifications). Over the course of the 27 days, participants spent 43.0 hours on average (SD = 22.1,10.3-121.8) using their smartphone, 2.9 hours (SD = 22.1,10.3-121.8) on locked devices. Overall, lock screens resulted in an average overhead of 2.9% (0.6- 9%).

Context and Satisfaction

Most (62.4%) device unlocks were sampled in private context. In 19.5% of the samples, participants indicated a semi-public setting, 18.2% of the environments were rated public. This is inline with previous work which found that mobile device interaction often takes place in private environments [121].

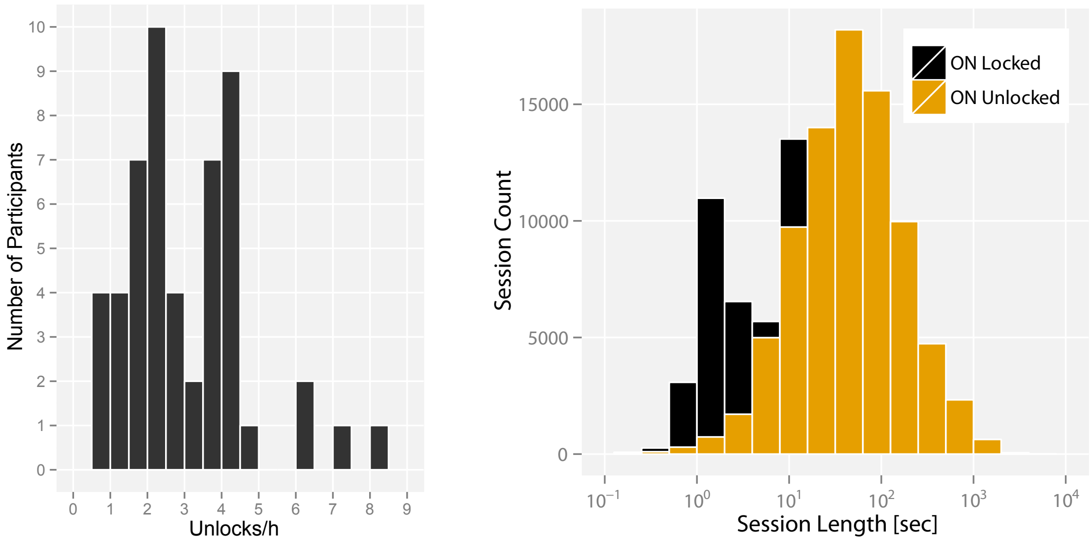


Figure 3.6: Unlock Frequency versus session length: Most sessions are very short.

Environment	# Situations	Slide-to-Unlock	Pin, Pattern	overall
private	2115 (62.0 %)	5.0 % (*sd* = 14.9 %)	32.7 % (*sd* = 36.0 %)	23.6 % (*sd* = 33.2 %)
semi-public	690 (20.2 %)	4.6 % (*sd* = 12.2 %)	23.0 % (*sd* = 29.3 %)	17.0 % (*sd* = 26.3 %)
public	605 (17.7 %)	6.2 % (*sd* = 20.1 %)	16.6 % (*sd* = 26.9 %)	13.2 % (*sd* = 25.2 %)
Overall	3410	5.3 % (*sd* = 15.8 %)	24.1 % (*sd* = 31.4 %)	17.9 % (*sd* = 28.6 %)

Table 3.1: Participants' dissatisfaction with their locking mechanisms by environment.

Within the data risk survey, participants rated the annoyance of the current unlock using a five-point Likert scale. Answers ranged from "not annoying at all" to "very annoying". Overall, participants were quite happy with most unlocks as 65.8% of all unlocks were rated "not annoying" or "not annoying at all". Nevertheless, 24.7% of the sampled unlocks were rated "annoying" or "very annoying". An analysis on user level reveals that only 12 of 52 participants indicated to be annoyed in more than 50% of the sampled situations. Figure 3.7 illustrates the answers grouped by unlock mechanism: While Slide-to-Unlock users rated 92.9% of all unlocks "not annoying" or "not annoying at all", only 56.0% of the pattern unlocks and only 49.7% of the PIN unlocks were rated the same. In contrast one third of all unlocks using PIN (33.1%) or Pattern (33.4%) were rated annoying. Using Slide-to-Unlock, only 5.5% of the unlocks were rated this way. An analysis considering usage frequency and the unlock context showed no clear trends.

In addition, we assessed satisfaction by asking Slide-to-Unlock users if they would have rather wished to have a secure lock screen in the current situation. Vice versa, we asked pattern and PIN users if a Slide-to-Unlock screen would have been desired. Table 3.1 illustrates the results which were gathered using five-point Likert scales clustered by environment. We

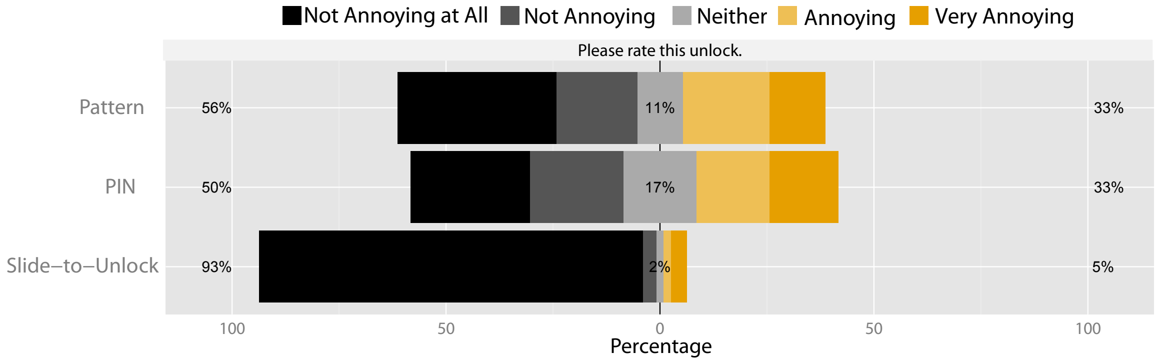


Figure 3.7: Unlock annoyance grouped by system (indicated using 5-point scales).

found that PIN and pattern users were more often dissatisfied with their current configuration while Slide-to-Unlock users indicated less dissatisfaction. While the analysis of unlock frequency and data sensitivity showed no trend, unlock context seems to influence satisfaction. The results indicate that code screen user (PIN, patterns) wished to have Slide-to-Unlock especially in private spaces and were more often satisfied when unlocking in public. In contrast, Slide-to-Unlock users were mostly satisfied with the status quo, even in public.

Data Sensitivity and Risk Perception

With the unlock risk survey, we asked participants to indicate the sensitivity of the data they are going to access. The answers were collected based on a five point scale ranging from "not sensitive at all" to "very sensitive". In 20.1% of the cases participants were not able to assess the sensitivity of the data and pressed a "do not know" button. Figure 3.8 illustrates the cases in which data sensitivity could be rated. For each user, we mapped the proportion of sensitive data accesses across the average time spent for unlocking each day. Overall, participants indicated in 25.3% of all sampled situations that the accessed data was "sensitive" or "very sensitive". Figure 3.8 shows that only ten (19.2%) users accessed sensitive data in more than 50% of their (sampled) unlocks. Additionally, we find interesting individual cases: The participant who spent most time for unlocking the device (highest value on the y-axis) indicated to access sensitive data in less than every fifth unlock. In addition, we find two Slide-to-Unlock users who indicated to access sensitive data in more than 50% of all cases. In the debriefing, we additionally asked participants to generally rate the sensitivity of their device's data. While almost half (48.6%) of the PIN and pattern users considered the stored data sensitive, only a quarter (23.5%) of users without a code-lock stated the same.

In addition, we investigated perceived risk concerning observation attacks and unwanted access. Therefore, we asked participants to indicate if the risk was present and if the potential attacker was known or unknown. Whenever the risk was present, participants additionally assessed the likelihood of an attack and indicated the severity if the attack was actually performed. Table 3.2 gives an overview of the results.

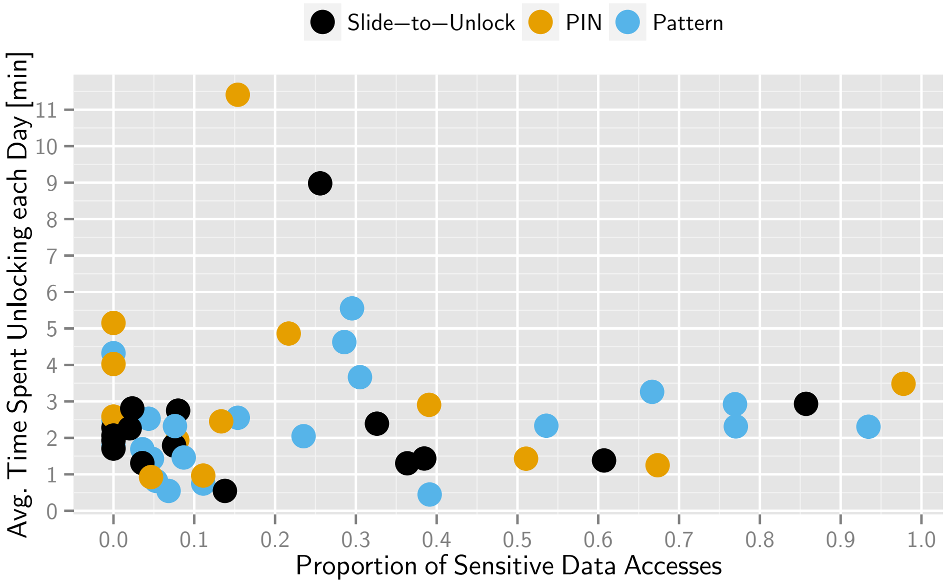


Figure 3.8: Accessed sensitive data by unlock system (indicated using 5-point scales).

The risk of unwanted access was reported in 7.7% of all sampled situations. While eleven participants did not indicate any risky situation, the individual count of participants who reported such risk ranged from one to 20 occasions. Table 3.2 shows that unwanted access was most often reported in private environments and potential attackers were known to the participants in most cases. Over 90% of all indicated occasions were reported unlikely. In addition, participants did hardly expect severe consequences. However, participants tended to rate unwanted access more likely and more severe when risk was perceived in public. During debriefing, some participants mentioned that probably not all cases had been sampled but confirmed that the sampled proportion matched their risk perception.

Observation risk was perceived more often than the risk of unwanted access. In 17% of all sampled situations participants reported that someone could have had "a view on the contents of [the] screen". Table 3.2 shows that this was most often possible (68.7%) for known people. Naturally, the proportion of unknown observers increases when unlocks are performed in more public environments. Overall, participants rated observations likely in 40.8% of all sampled situations. However, the severity of such observation events was rated low in most cases, ranging from 68.3% in public settings to 92.9% in private environments. Only 11 of the 3410 (0.3%) sampled situations were perceived likely and “severe” or “very severe” at the same time. Seven (63.6%) of these critical situations were reported in public spaces. Furthermore, additional countermeasures (e.g., covering the screen) were reported in only 52 (2.8%) sampled situations. Again, participants mentioned during debriefing that probably not all situations were sampled but agreed that the sampled data was representative.

Environment	# Situations	Known	Unknown	Unlikely	Low Severity
Unwanted Access					
private	131 (53.5 %)	97.7 % (128)	2.3 % (3)	92.4 % (121)	86.3 % (113)
semi-public	75 (30.6 %)	70.7 % (53)	29.3 % (22)	93.4 % (70)	64.0 % (48)
public	39 (15.9 %)	23.1 % (9)	76.9 % (30)	79.5 % (31)	18.2 % (11)
Overall	245 (7.7 %)	77.6 % (190)	22.4 % (55)	90.6 % (222)	70.2 % (172)
Shoulder Surfing					
private	182 (31.5 %)	99.5 % (181)	0.0 % (1)	56.6 % (103)	92.9 % (169)
semi-public	185 (31.0 %)	82.7 % (153)	17.3 % (32)	65.4 % (121)	84.9 % (157)
public	211 (36.5 %)	29.9 % (63)	70.1 % (148)	56.0 % (118)	68.3 % (144)
Overall	578 (17.0 %)	68.7 % (397)	31.3 % (181)	59.2 % (342)	81.3 % (470)

Table 3.2: Unwanted access and observation occasions by environments and potential attackers. The last two columns give percentages of likelihood and severity of consequences.

3.2.4 Discussion and Implications

In this Section, we summarize the main findings and discuss the implications for feasible unlock mechanisms for mobile devices.

Many Short Periods of Use Create a High Authentication Overhead

The comparison of the assessed unlock frequency and the measured data revealed that many users tend to underestimate their daily unlock numbers. This indicates that current solutions support subliminal interaction and that the perceived effort is often kept adequately low. Concerning the speed of single unlock events, “adequately” would stand for unlock times between 2.7 seconds (Slide to Unlock) and 4.7 seconds (PIN). At the same time, we found that the common unlock behavior comprises a high number of unlocks combined with relatively short periods of use. Over the course of 27 days, current solutions generated an average overhead of 1.2 hours (2.9%) for all users and up to 5.1 hours (9%) for heavy users.

Such numbers show that every additional second has a high impact on the overall unlock effort. However, this is often neglected as novel authentication concepts are usually evaluated in lab experiments and in the wild unlock behavior is seldom considered. Chapter 2 revealed that most alternative authentication concepts demand authentication times of ten seconds and more. While ten seconds might be rated acceptable in short-term lab tests, actual unlock behavior would translate this time to a monthly additional effort of several hours. Considering the short periods of use, this would easily introduce authentication overheads over 10%, probably more than the average user is willing to invest. We therefore conclude that feasible unlock mechanisms must consider actual usage patterns and thus allow fast interaction which is comparable to PIN and patterns.

Sensitive Data is Seldom Accessed

According to our online survey, most participants motivate the use of secure unlock screens by general protection goals: “access control” was most often mentioned. In addition, the goal to “protect information” was frequently stated. In practice, we observed that only a quarter (25.3%) of all sampled unlocks was actually performed to access sensitive data. In most cases, participants stated they would interact with non-sensitive data which presumably not required protection. This indicates that only a subset of the demanded unlocks is actually justified by the protection goals. As a consequence, users might judge a good portion of the unlock effort unnecessary and rather decide to completely forgo secure unlock screens. Indeed, “inconvenience” was often stated as a reason for not using secure lock screens. At the same time, we found that the proportion of accessed sensitive data is not a predictor for the use of secure lock mechanisms. Several participants frequently accessed sensitive data but still relied on the insecure Slide-to-Unlock mechanism.

We conclude that even if the all-or-nothing lock paradigm is not a main factor for the adoption of secure concepts, it may indeed prevent some users from using such systems. We argue that content-dependent security models (e.g., on application level) could decrease inconvenience caused by the authentication overhead and would simultaneously couple the additional effort with users’ primary protection goals. Finally, reducing the number of time-consuming unlocks would probably increase the general acceptance of secure lock screens.

Often Discussed Risks are Seldom Perceived

Secure unlock mechanisms shall prevent unwanted access. However, the risk of unwanted access was perceived very low. According to the participants’ perception, it was possible in only 8% of all sampled situations. In addition, most reported cases were rated unlikely and most users would not expect severe consequences. Nevertheless, in public spaces participants rated unwanted access more likely and tended to expect more severe consequences. Similarly, observation resistance is an often discussed requirement for mobile authentication concepts. Chapter 2 showed that system designers often accept decreased efficiency to achieve this goal. However, we found that observation risks are also seldom perceived. Even if observation risks were perceived possible more often (17%) than the risk of unwanted access, high risks with severe consequences were reported very seldom (0.3%).

We argue that the uncritical perception of risks does not suggest a general absence of risks. As a consequence, the results neither indicate that secure lock mechanisms are dispensable nor do they generally question the importance of observation resistance. However, we assume that users are not willing to invest additional effort to be protected from risks which are rarely perceived critical. Therefore, increased practical security is hardly a sales argument for novel authentication mechanisms and the additional effort a user is willing to invest needs to be carefully assessed. As we cannot assume permanent high risk situations, secure authentication mechanisms must be comparably fast to current concepts or demand extra effort only if specific risks are actually present or perceived.

Current Solutions are Sometimes Annoying and Seldom Satisfying

In the online survey, most participants attested that current unlock mechanisms are easy to use. Nevertheless, every second respondent agreed that they can be annoying sometimes and a quarter of the participants still desired easier mechanisms. In the field, we found interesting differences between Slide-to-Unlock users and code lock users. While only 5% of the Slide-to-Unlock events was rated annoying, both PIN and pattern users were annoyed by 33% of the sampled unlocks. This indicates that secure unlock screens already demand a noticeable extra cost compared to Slide-to-Unlock. Furthermore, we found that Slide-to-Unlock users rarely felt the need for secure lock screens, while lock screen users were dissatisfied with their settings in almost a quarter of all situations. This was particularly the case when serious threats were perceived unlikely (e.g., in private spaces). In addition, we found that many users felt protected, even when no secure lock screen was used. Such users relied on physical countermeasures like concealing the smartphone or leaving it at a safe place.

The results show that there is a very thin line between satisfaction and annoyance. While the measured difference between unlock patterns and Slide-to-Unlock was marginal (0.3 seconds), the effects on subjective ratings were strong. We assume that the dissatisfaction of code users was mainly provoked by the knowledge of a faster, albeit insecure, alternative: Slide-to-Unlock. Similarly, physical protection in risky situations is seen as a feasible alternative to permanently using code locks. This confirms that users permanently evaluate the perceived benefits of a system and put them into the context of existing alternatives. As a consequence, we conclude that novel authentication systems need to be as satisfying as current unlock mechanisms to get accepted. To satisfy Slide-to-Unlock users, secure concepts may need to perform even better than PIN or patterns. Context-dependent security mechanisms which adapt the security level (and extra costs) to the current situation may be one promising direction. However, “context” is hard to define and cannot be simplified as location information (e.g., public).

3.2.5 Limitations

Even if the online survey and the field study were thoroughly designed and conducted, they have inherent limitations which need to be addressed.

The results of the online survey are based on self-reported data and should therefore only be interpreted as an indicator for real-world behavior. The field study achieved higher external validity by sampling user behavior and risk perception in-situ. However, it is important to note that neither the number of samples per participant nor the proportion of lock screen types or environments was counterbalanced. Nevertheless, we assume that the differences between the different sample sizes are not crucial as high numbers were collected for every combination. While participants confirmed that the results of the experience sampling represented everyday life, we like to mention that some situations might still be underrepresented. This might especially be true, whenever situations were very short (e.g., unlocks in an elevator). In addition, we assume that participants used the “not now” button primarily in stressful situations where answering a questionnaire was not feasible. As a consequence, such critical situations might be underrepresented as well. Still, we are confident that the data is valid and representative as such situations were not reported during debriefing.

The experience sampling itself might have an impact on unlock behavior and risk perception which would influence the results. In the debriefing, participants were asked to report potential study impact. Indeed, four participants reported a minor impact on their behavior. Three participants mentioned that they might have used the device less frequently, especially in the beginning of the study. Another participant increased the auto lock interval from 30 to 90 seconds. In addition, ten participants reported that taking part in the user study made them more aware of their own unlock behavior without actually changing the behavior. We argue that such reports indicate rather minor impact and are therefore confident that real-life behavior was not significantly influenced.

Finally, we tested a convenience sample which was mainly based on students. As a consequence, participants were higher educated and younger than the average population and the results cannot be generalized to any specific population. However, we argue that students were well suited for a first evaluation as they are very likely to experience a large set of diverse situations ranging from office settings to uncontrolled public spaces (e.g., concerts).

3.2.6 Summary

In this Section, we presented the results of an online survey and a longitudinal field study. The online survey provided an overview over users’ attitudes towards authentication on mobile devices as well as alternative protection strategies. The field study which combined traditional activity logging with unobtrusive experience sampling allowed gathering in depth insights into real-life unlock behavior and risk perception.

We learned that the decision to use or to not use secure lock mechanisms is influenced by various factors. Non-users justified their behavior mainly by an absence of threats. But inconvenience played an important role, too. Indeed, we showed that users check their devices frequently and periods of use are usually short. As a result, high authentication overheads were measured even with the currently used unlock mechanisms. In addition, we showed that sensitive data is seldom accessed and that most unlocks take place in private spaces. Moreover, since common risks were seldom perceived and mostly rated unlikely, it was not surprising that a good portion of the code-based unlocks was rated annoying.

This Section did not focus on the peculiarities of gesture-based authentication but presented important requirements for all feasible authentication mechanisms on mobile devices. Two insights are particularly important for the remainder of this work. First, authentication methods must be very fast as every additional second has a high impact on the overall overhead. Secondly, since risks are seldom perceived, users are most probably not willing to invest additional effort for improved security.

3.3 Usability of Grid-based Gestures in the Wild

The previous Section analyzed user attitudes and user behavior in the wild and defined important requirements which apply to all knowledge-based unlock mechanisms. In this Section, we focus on the specialties of gesture-based authentication and assess how its usability compares to the traditional PIN-based approach. Chapter 2 indicated that graphical gesture-based authentication mechanisms provide superior usability and are especially well-suited for mobile devices. However, the literature review did hardly provide evidence that this actually holds true in the user's real life. To fill this gap of knowledge, we now compare the two most prominent knowledge-based unlock mechanisms in a field study: Grid-based gestures and PIN. We shed light on the following main *research questions*.

RQ1 How *efficient* are grid-based gestures and how is efficiency perceived?

RQ2 How *effective* are grid-based gestures and how is effectiveness perceived?

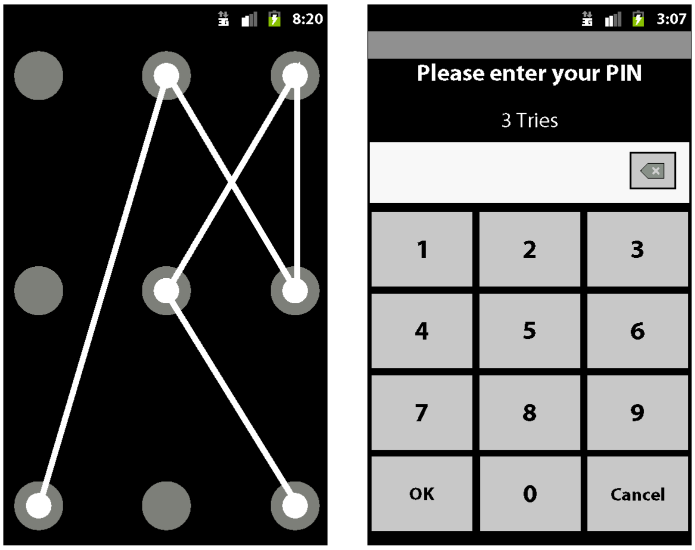
RQ3 How *memorable* are grid-based gestures and how are they learned?

RQ4 How *satisfying* are grid-based gestures and what are the influencing factors?

The presented results are based on a 21 days between-groups field study including 34 pattern users and 26 PIN users. We increased internal validity by following a strictly structured procedure. For example, PINs and patterns were controlled in length and complexity and all participants contributed the same number of data points. Even if this means that external validity was reduced, the study design allowed an accurate assessment of both authentication concepts in the real world.

This Section will show that PIN *performs* better when looking at quantitative performance data but gestures are *perceived* equally good (*RQ1*). Even if more errors were logged, unlock gestures were *perceived* easy to use and error recovery was rated better (*RQ2*). The detailed analysis of the observed errors revealed interesting insights into the relationship between error prevention and error recovery. In addition, a novel taxonomy for gesture-based errors will be presented. While gestures and PINs seem equally *memorable* (*RQ3*), participants were generally more in *favor* of the gesture-based approach (*RQ4*). Overall, this Section contributes by providing real-world evidence for the assumption that grid-based gestures are a feasible unlock mechanism for touch-based mobile devices and a promising alternative to (alpha)numeric solutions.

This Section is partly based on von Zezschwitz, E., Dunphy, P., & De Luca, A. (2013, August). Patterns in the wild: a field study of the usability of pattern and pin-based authentication on mobile devices. In Proceedings of MobileHCI'13. [266]. Please refer to the beginning of this thesis for a detailed statement of collaboration.



Please enter your PIN

3 Tries

Figure 3.9: The two prototypes of the user study. The left image shows the interface of the gesture-based system. On the right side, the PIN system is illustrated.

3.3.1 Research Context and Motivation

Chapter 2 provides an in-depth discussion of grid-based authentication concepts. The discussion indicates that grid-based gestures are easy to recall and well-suited for mobile devices. While all presented concepts were specifically designed for mobile devices, Android unlock patterns have become the first gesture-based system which was actually widely accepted [104]. In September 2015, 1.4 billion devices supported grid-based gestures[9]. At the same time, none of the prior studies has evaluated such concepts over longer periods of time and field performance has not been investigated systematically. Therefore, this Section provides the results of the first empirical longitudinal study on the performance and the likeability of gesture-based authentication in the wild.

3.3.2 User Study

The user study had the primary goal to collect performance data and user feedback over a longer period of time and outside of the laboratory environment. The PIN group served as a baseline and was added for comparison. Due to organizational reasons, the PIN group started after the gesture group was finished.

Prototypes

For each system, a distinct prototype was developed. The prototypes which are illustrated in Figure 3.9 were implemented as standalone applications for Android 2.1 or higher. Both systems represented common graphical user interfaces which can be found on current mobile devices. To prevent graphical bias, the design was kept very simple. As customary, PINs had to be explicitly confirmed using the “OK” button. In addition, input could be corrected using “backspace” or “cancel”. The gesture system was inspired by current Android implementations. Therefore, input was confirmed implicitly by lifting the finger and could not be corrected. However, in contrast to Android standards, our implementation allowed to revisit and to skip dots. Therefore, the prototype supported more complex gestures and allowed to examine the consequences of providing a larger theoretical patterns space.

Both prototypes provided two distinct modes: a training mode allowed unlimited interaction and helped participants in getting used to the interface. The study mode requested 21 authentications with a maximum of one authentication per day. Each authentication session could comprise a maximum of three attempts. After three failed attempts or after a successful unlock, the system was disabled until the start of the next calendar day. In study mode, user interaction was constantly logged. This included single touches as well as the input of gestures (as a sequence of cell activations) or PINs.

Method

The user study was designed following a between groups longitudinal design. Since we opted for a controlled setting, we assigned PINs and patterns and fixed the number of samples. In the gesture group, we randomly assigned Android-conform and more complex patterns. After one day of training, the performance test started. Over the course of 21 days, participants were asked to authenticate once a day using their assigned secret. After the performance test, we collected qualitative feedback using a questionnaire. After further 14 days had passed, participants were spontaneously asked to recall their secret (memory test).

Password Assignment By assigning secrets, we had full control over length and complexity which was important for comparability. PIN users were assigned four-digit PINs which represents a common length on mobile devices. Using four digits results in a theoretical password space of 10,000 (10^4) secrets. The pattern group was assigned gestures which comprised six cells. Considering Android restrictions this would result in a comparable pattern space of 26,016 gestures[10]. However, allowing more complex gestures increases this number to 294,912 ($9 * 8^5$). That is, the starting point can be freely chosen from nine cells and subsequent moves can reach eight cells (nine minus the current position).

[9] http://www.androidcentral.com/google-says-there-are-now-14-billion-active-android-devices-worldwide – last accessed: 2015/11/26.

[10] https://www.quora.com/How-many-combinations-does-Android-9-point-unlock-have – last accessed: 2015/11/28

Installation and Training On the first study day, participants received an email containing a download link for the prototype, an installation instruction, a unique secret (PIN or gesture) and an anonymous user id. After downloading and installing, participants were asked to enter their user id. The user id determined the assigned secret and was used to relate the logged performance data to the later questionnaire. Next, participants were encouraged to play with the prototype and to repeatedly authenticate using the assigned secret (training mode). Whenever participants felt ready, they could stop the training and the user study began. However, the training stopped automatically after one calendar day.

Performance Study In the following 21 days, participants had to authenticate once per calendar day. As already mentioned, we allowed a maximum of three attempts. That is, a session finished with a successful authentication or after three failed attempts. E-Mail reminders were sent once a day. Such mails did not contain personal data, especially the secret was not provided. If participants still forgot to authenticate, we allowed an extension of one day. However, a maximum of five additional days was granted.

Debriefing After the performance test, participants were invited to the lab for debriefing. We collected the log-data from the participants' devices and helped with uninstalling the application. Next, participants filled in a questionnaire which collected qualitative feedback concerning usability and likeability. In exceptional cases, when participants were not able to appear in person, the debriefing was performed remotely using video chat software.

Memory Test During the debriefing, we mentioned that we might contact participants in the future if we had more questions. After 14 days had passed, we arranged a spontaneous meeting and asked participants to recall their secret (PIN or pattern) using a printed version of the prototype. We opted for a printed version of the user interface as it is independent from specific form factors. Again, participants had three attempts. After the recall test, the memorability of the respective system was rated using Likert scales.

Participants

Participants were recruited using the university mailing list and social networks. All participants were required to own a smartphone with Android 2.1 or higher. For organizational reasons, the PIN group was recruited after the pattern group had finished. As an incentive, participants had the chance to win one of two gaming consoles.

The pattern group started with 38 participants of whom 29 finished both the performance test and the recall test. The average age of the valid 29 participants was 26 years (SD = 4, 19-36). Eleven participants were female, 18 male. 21% stated to use unlock patterns in daily life to authenticate on the smartphone. The PIN group started with 30 participants out of which 24 contributed valid data sets. The average age of those 24 participants was 27 years (SD = 4, 21-42). Seven of them were female, 17 male. 46% stated to use PIN on their smartphone to authenticate. Participants of both groups were highly educated as 93% of the pattern group and 92% of the PIN group held a university-entrance diploma.

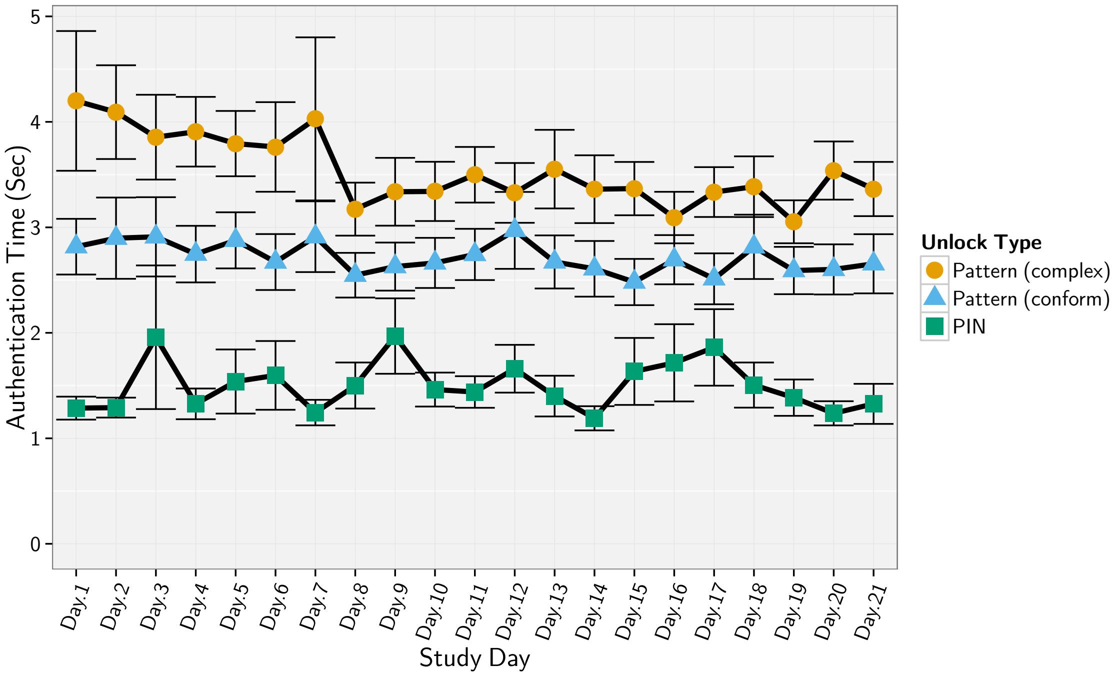


Figure 3.10: The measured authentication times of PIN, Android-conform gestures and more complex gestures mapped on the 21 days of the user study. PIN performed best, followed by Android rules conform patterns and complex patterns.

Results

We focus on a general comparison of PINs and patterns. Furthermore, we compare 13 Android-conform patterns to 16 patterns, where dots are skipped or visited several times, to examine if the larger password space comes with usability drawbacks.

Efficiency

Efficiency is analyzed based on the average input times of successfully finished authentications. Analyzing failed attempts would skew the results: For example, the set of gesture errors comprises deliberately aborted attempts which are usually very short. For the gesture scheme, time measurement started with the first touch event and stopped as soon as the finger was lifted. For PIN, we started the time measurement with the activation of the first button and finished after the last digit (button) was released. In contrast to the gesture mechanism, the PIN system required explicit confirmation of the input. However, we excluded this step from the analysis to get comparable input data.

Figure 3.10 illustrates the average authentication times of PIN and pattern users, mapped on the 21 days of the user study. The data of pattern users was split according to the complexity of the used gesture. To investigate the overall performance, we conducted a one-way independent analysis of variance comparing the 21-days average of PIN and both pattern groups. The results reveal that the used authentication token has a significant impact on

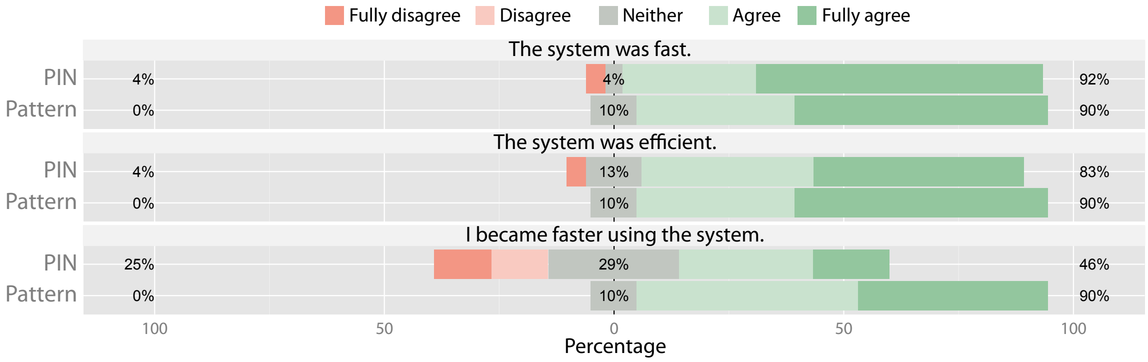


Figure 3.11: Users' opinions towards the efficiency of the tested systems. In contrast to the measured data, the pattern system was not perceived slower than PIN.

authentication time, $F_{2,52} = 30.66$, $p < 0.001$. Bonferroni corrected post-hoc tests reveal that PIN (1501 ms, SE = 165, 844-3141) users performed significantly faster than pattern users, $p < 0.001$. In addition, using an Android-conform gesture (2714 ms, SE = 216, 1315-4655) resulted in significantly faster authentication times than using a more complex gesture (3531 ms, SE = 208, 1618-5364), $p < 0.05$.

To analyze if users became faster over time, we performed a one way repeated measures ANOVA. The results confirm the significant main effect of the authentication token, $F_{2,40} = 19.94$, $p < 0.001$. However, no significant effect was found for the study days, $F_{6.33,253.37} = 0.73$, $p > 0.05$, Greenhouse-Geisser corrected ($\varepsilon = 0.32$). Furthermore, no interaction effect of $token * days$ was found, $F_{12.67,253.37} = 0.73$, $p > 0.05$. Bonferroni corrected pair-wise comparisons of the authentication token revealed a significant mean difference between both PIN and complex patterns (1959 ms, SE = 321) and PIN and conform patterns (1169 ms, SE = 321), $p < 0.001$. However, the observed mean difference between conform and complex patterns was not significantly different (789 ms, SE = 368), $p = 0.11$.

In contrast to the pattern concept, PIN allowed undo operations. Over the course of 21 days, we observed 20 corrected attempts (using backspace) and five cancelled attempts (using the cancel button). Comparing such authentication attempts (Mn = 5720 ms, SE = 665) which included undo actions with the other successful attempts (M = 1314 ms, SE = 33) reveals that such undo operations significantly slow down the input process, $t_{24.12} = -6.62$, $p < 0.001$, $r = 0.80$.

In addition to the quantitative data, we assessed the perceived efficiency using five point Likert scales ranging from "fully agree" to "fully disagree". Figure 3.11 illustrates the data. Since a Wilcoxon rank-sum test showed no significant impact of pattern complexity ($p > 0.05$) the data was combined and grouped by authentication system. Overall, 21 (92%) PIN users and 26 (90%) pattern user agreed that authentication speed was fast. When asked for efficiency, 20 (83%) PIN users and 26 (90%) pattern users attested positive values. While these differences were not significant ($p > 0.05$), we found that significantly more

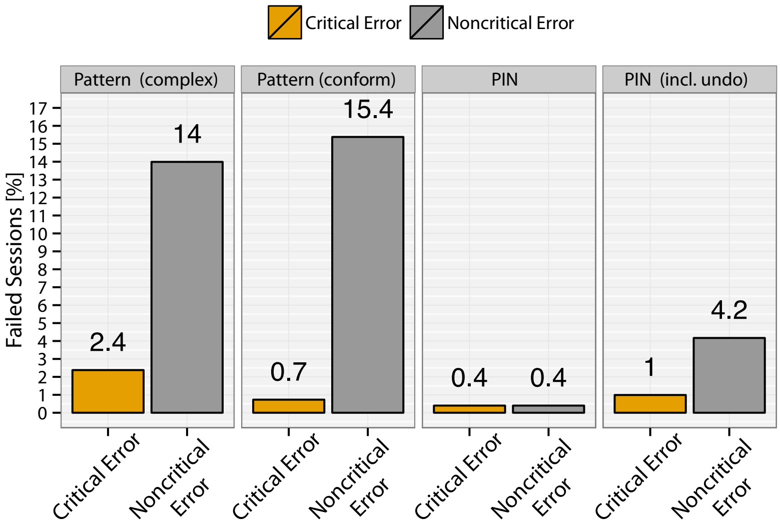


Figure 3.12: The number of errors subdivided by the complexity of the used gesture (pattern only) and by the use of undo operations (PIN only).

participants using the pattern scheme (90%) assessed their authentication speed to become faster over the three weeks period, $W_s = 472.50$, $z = -3.30$, $p < 0.001$, $r = -0.45$. In the PIN group, only 46% stated the same. The results indicate that the measured differences were not reflected by the users' opinions as both systems were perceived fast and efficient to use. In addition, even if the quantitative analysis did not reveal significant changes over time, most pattern users reported training effects.

Effectiveness

Effectiveness is analyzed based on the number of errors. In addition, we assess the sources of errors and the perceived ease of use. Android security usually allows up to five failed attempts before the system will be locked for a certain period of time (usually 30 seconds). However, the security configuration can vary between different device classes and between different manufacturers. As a consequence, we decided to categorize errors more conservative and compare them to the stricter but standardized security policies of automated teller machines (ATMs) [134]. ATMs allow a maximum of three failed attempts before the used authentication token (i.e., PIN) becomes ineffective. After such an error, some form of fallback authentication is required to reactivate the account. Therefore, we define three consecutive failed attempts as *critical error*. Consequently, one or two failed attempts are categorized as *non-critical error*.

The analysis is based on 504 (21 days $*$ 24 users) PIN-sessions and 609 (21 days $*$ 29 users) pattern-sessions. In 273 (13 users) pattern sessions an Android-conform gesture was used, the remaining 336 (16 users) sessions were conducted with more complex gestures. Each authentication session can include one, two or three attempts.

An independent t-test comparing the overall error rate of PIN and patterns reveals that PIN was significantly less error-prone than gesture-based authentication, $t_{29.93} = -6.26$, $p < 0.001$, $r = 0.75$. Overall, we observed only two critical (0.4%) and two non-critical (0.4%) PIN errors. In contrast, gesture-based authentication led to non-critical errors in 89 sessions (14.6%) and to ten (1.6%) critical errors. However, error rate was not significantly influenced by complexity ($p > 0.05$). Users with Android-conform gestures failed non-critically in 15.4% of the sessions while the use of more complex gestures led to non-critical errors in 14.0% of the sessions. In addition, 2.4% of the sessions comprised a critical error when more complex gestures were used and two (0.73 %) sessions failed critically using Android-conform gestures. In contrast to the pattern group, PIN users were able to correct input errors using the provided undo operations. As this aspect reduced the resulting error rate of the PIN prototype, we added all correct-ed sessions to the data set and performed a second analysis. The results indicate that, even without undo operations, PIN users performed significantly better, $t_{51} = 2.81$, $p < 0.05$, $r = 0.37$. Figure 3.12 compares the described error rates considering gesture complexity and the use of undo operations.

To further investigate why more people failed when authenticating with gestures, we had a closer look into the characteristics of the logged attempts and developed a novel taxonomy for the categorization of gesture-based errors. Since the scheme was primarily based on a thorough analysis of logged data and theoretical assumptions, we additionally presented the results of the categorization to four randomly selected participants. Participants informally confirmed that the assumed reasons matched their real reasons for failed authentications. Overall, we defined five reasons for slip errors and three reasons for memory-related failures. As defined in Section 3.1.1, slip errors are the consequence of inaccurate input, while memory-related failures are the consequence of inaccurate recall. The taxonomy will be illustrated based on the pattern “ ” (see Figure 3.13, right).

Slips are described by the following stroke types:

Aborted A subset of the correct cells in the correct order (Figure 3.13,1).

Additional An additional stroke which is not part of the correct pattern (Figure 3.13,2).

Distributed A correct pattern which is distributed over multiple attempts (Figure 3.13,3).

Missing A missing stroke that is not at the end of the pattern (Figure 3.13,4).

Close A wrong stroke which connects a wrong cell *close* to the correct cell (Figure 3.13,5).

Memory-related errors are described by the following stroke types:

Repeated A repetition of the same wrong pattern (Figure 3.13,6).

Distant Wrong strokes that are not direct neighbors of the correct stroke (Figure 3.13,7).

Mirrored Wrong strokes that are mirrored versions of the correct strokes (Figure 3.13,8).

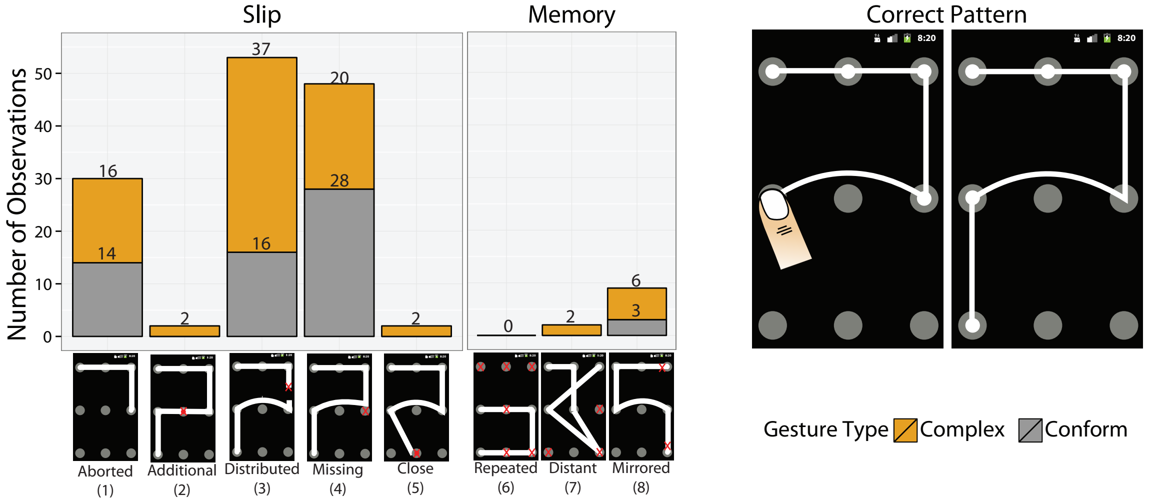


Figure 3.13: The assessed origin of gesture-based authentication errors. The source of error was assessed based on a novel taxonomy. According to our analysis, most observed errors were based on slips. Memorability issues played a minor role.

Based on the taxonomy, we categorized 135 (93%) slips and 11 (7.5%) memory-related errors. Whenever an error showed characteristics of both a slip and a memory-related failure, we counted it as a memory-related problem. For example, we observed attempts which were aborted after a mirrored (sub)pattern. We assumed that such mirrored strokes are hardly performed due to inaccurate input. Figure 3.13 gives an overview of the observed sources of errors. The analysis revealed that 58 (43%) slips and three (27%) memory-related errors were performed with Android-conform gestures.

As indicated by Figure 3.13, most slips (53) led to distributed gestures which were cut in multiple parts and thus were logged as separate attempts. 16 (30%) of these were based on conform gestures. Those errors are the consequence of short interruptions which occur whenever the user briefly lifts her finger from the display. Since the system has no explicit confirmation mechanism, each interruption is logged as a (failed) authentication attempt followed by another (failed) attempt. The second largest group is based on missing strokes (36%), 28 (58%) of these based on conform patterns. Furthermore, we observed 30 (22%) aborted attempts, 14 (47%) of these with conform gestures. Within the "aborted" group, we found wrong and correct stroke sequences. Similarly to distributed patterns, the abort of a correct pattern seems to be based on unintended interruptions (slips). In contrast to distributed gestures, the interruption was probably recognized by the user. Furthermore, we assume that the abort of a wrong pattern sequence is the consequence of a recognized error and therefore represents an intended action and not an error itself. One frequently observed error, which also led to the abort of many attempts, was accidentally touching a cell which actually should have been skipped. Finally, we found that all but one critical error had been a consequence of inaccurate input (slips).

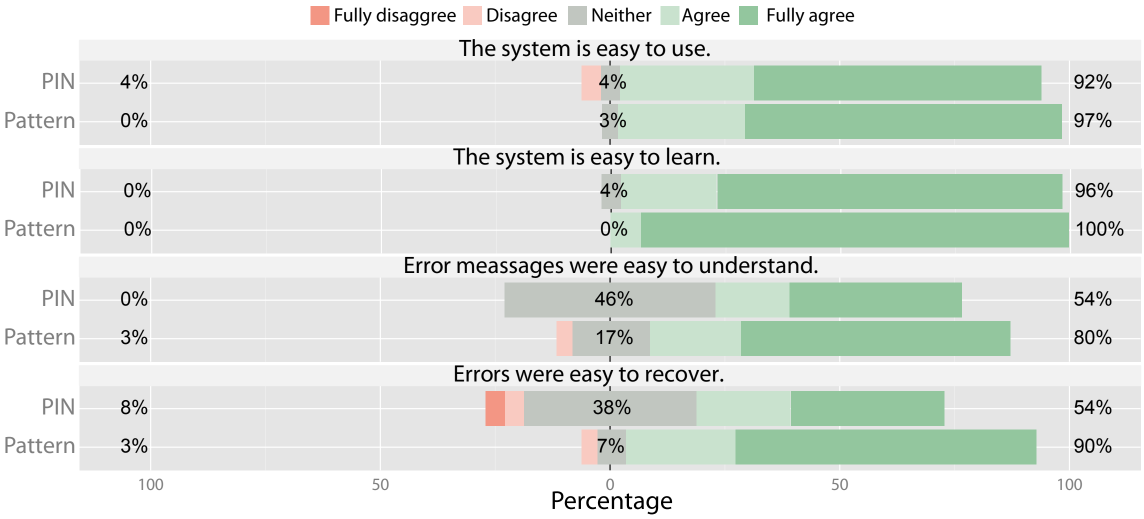


Figure 3.14: Users' opinions towards ease of use. Both systems were rated easy to learn and easy to use. In addition, pattern users were more positive about error recovery than PIN users.

In the PIN group, where six failed attempts were logged, one critical error was based on inaccurate recall: The user entered three wrong PINs which were all based on the transposed correct PIN. The second critical error was based on a hardware problem and thus does not fit in any category. The non-critical failures were based on missing digits (slips).

The qualitative feedback is illustrated in Figure 3.14. The results indicate that gesture users were not annoyed by the number of errors. Indeed, both groups rated the used prototypes easy to use. This was the case for 27 (93%, Mdn = 5) pattern users and 21 PIN users (88%, Mdn = 4), $W_s = 595.50$, $z = -1.05$,$p > 0.05$, $r = -0.14$. Interestingly, 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. This indicates that quick recovery is more important for the perceived usability than error avoidance. Furthermore, 79% (Mdn = 5) of the pattern users and 54% of the PIN users (Mdn = 4) stated that error messages were easy to understand, $p > 0.05$. This is an unexpected rating of the gesture group, considering that 39% of all slips were based on distributed patterns which indicates that error messages were not recognized or ignored. Finally 96% (Mdn = 5) of the PIN users and all pattern users (100%, Mdn = 5) agreed that the system was easy to learn.

Likeability

According to the participants' performance ratings, gestures were perceived equally efficient and equally effective to PINs. This somewhat contradicts the quantitative assessment which indicated significant differences. This Section focuses on user experience and indicates that likeability is an important factor that also influences the usability perception of the systems.

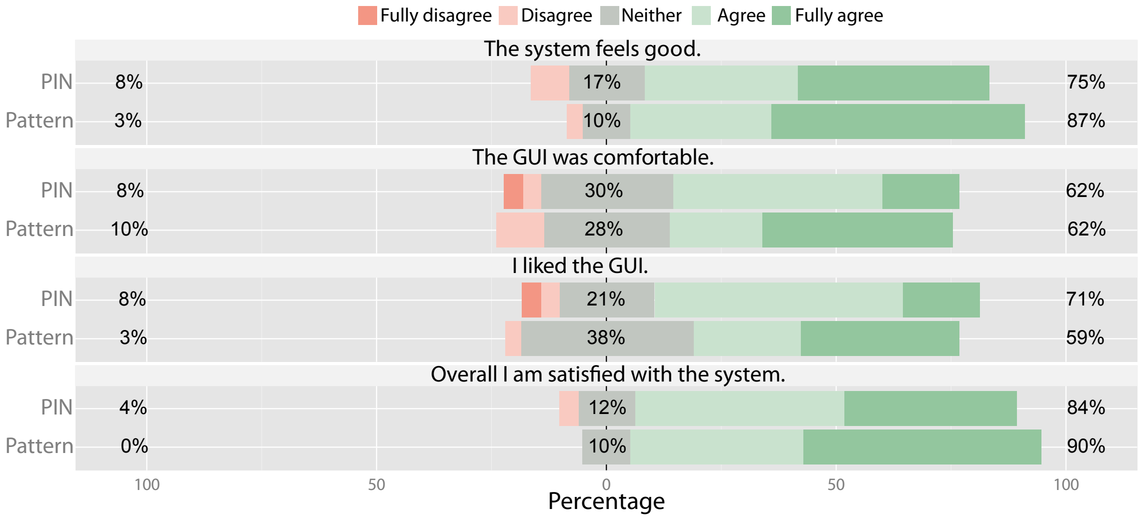


Figure 3.15: Users' likeability ratings. Most participants agreed that using the gesture-based authentication system feels good. In addition, gesture users indicated satisfaction.

As seen in Figure 3.15, 59% (Mdn = 4) of the pattern users and 71% of the PIN users (Mdn = 4) liked the graphical user interface, $p > 0.05$. This indicates that graphical attributes have not been the main influencing factor for the rather positive perception of the systems. Indeed, none of the GUIs was optimized for good appearance. We rather opted for reduced designs that minimized the effects of graphical differences. When we asked for interaction comfort, both prototypes were rated almost identical. Only 62% of the pattern users and 62% of the PIN users agreed that interaction was comfortable, $p > 0.05$. Therefore, the specific likeability ratings do neither indicate clear preferences nor do they indicate that one of the systems was particularly pleasant to use.

Nevertheless, when we asked for the overall likeability, ratings were very positive and skewed towards the pattern system. Most (86%, Mdn = 5) of the gesture users agreed that using the system felt good. In the PIN group, 75% of the PIN users (Mdn = 4) stated the same, $p > 0.05$. Finally, the overall likeability was confirmed by fact that 90% of the gesture group and 83% of the PIN group indicated to be satisfied with the tested system, $p > 0.05$.

Memorability

In the questionnaire, participants were asked to indicate the number of inputs which were required to memorize the assigned token. Twelve (41%) gesture users and 17 (70%) PIN users reported they had memorized their token after the first input. Only four of these twelve pattern users got assigned a complex gesture. Furthermore, 16 (55%) participants of the gesture group and six (25%) PIN users indicated they required two or three inputs. One PIN user and one pattern user (with a complex gesture) needed more than three inputs to learn the token.

The spontaneous recall test confirmed the good memorability of PIN and gestures. In the PIN group, 22 participants (92%) remembered their token, 19 (86%) among them needed only one of the three possible attempts. In the pattern group, 26 participants (90%) could recall the correct gesture, 23 (89%) among them within the first attempt. Two of the three participants who failed to recall their pattern had used a complex gesture. The participant with the Android-conform gesture remembered the correct shape but failed to recall the correct starting cell. While all users of the gesture group agreed that they learned their credentials on a motor or visual basis, 16 (67%) participants of the PIN group stated the same. They reported to recall the pattern, which evolves from connecting the intended buttons with imaginary lines. The rest of the PIN group associated previous knowledge or "learned it by heart". When we asked gesture users to compare their experiences to previous experiences with PINs, all agreed that gestures were equally easy to memorize.

Overall, we found no evidence for the assumption that the gesture-based approach is actually easier to recall than the numeric solution. Instead, good recall rates and steep learning curves of both systems support the assumption that memorability is indeed a minor problem as long as such concepts are frequently used.

3.3.3 Discussion and Implications

In this Section, we discuss the results of the longitudinal field study as well as the recall test and draw implications for the feasibility of gesture-based authentication.

PINs are Measurably Faster and Less Error-Prone

The 21-days performance test revealed that the system-type significantly influenced the success rate and the input speed of participants. We found that the PIN system allowed a more efficient and a more effective authentication on mobile devices. Overall, users of the gesture system needed more than twice as much time and made significantly more errors. While gesture complexity did not influence error rates, the more complicated interactions of non-conform patterns measurably slowed down the process.

At the first glance, the results seem to contradict previous findings reported in Section 3.2 which indicated that pattern users authenticate faster than PIN users. However, in contrast to the previous study, the measurement was not biased by other user actions (e.g., viewing notifications). Therefore, we argue that the data presented in this Section represents a more precise assessment. The differences may be partly explained by the fact that PIN-based concepts are widely-used and thus participants were more trained. At the same time, touching six cells may generally take more time than entering four digits. However, Android users need to select more cells to achieve a similar theoretical security to PINs. We conclude that considering only quantitative performance data, grid-based authentication is neither more efficient nor more effective than PIN-based authentication. However, according to the lessons learned from Section 3.2, the performance of the gesture system is still within an acceptable range for mobile applications.

Grid-based Gestures are Perceived Equally Fast and Easy to Use

In contrast to the measured differences, the usability ratings indicated no superiority of PINs. Both systems were rated fast and easy to use. In addition to the perceived performance aspects, we found that users overall liked the systems. Especially the gesture-based concept was reported to trigger good feelings and most users were very satisfied.

Therefore, we conclude that despite being measurably slower and more error-prone, the gesture-based system provided acceptable efficiency and effectiveness. At the same time, users were probably willing to accept the slower input as they overall liked the system. That is, the reduced performance might have been partly counterbalanced by good user experience. For example, the continuous movements of grid-based gestures might be more comfortable than multiple interrupted interactions (i.e., PIN-entry). In addition, fast error handling had a positive effect on the perceived ease of use. The next Section will discuss this aspect in detail. In summary, the results show that qualitative and quantitative data needs to be analyzed to assess the performance of a system. While measured performance is a predictor for perceived performance, user ratings may be influenced by various additional factors (e.g., likeability).

There is a Trade-off between Error Recovery and Error Prevention

The results show that the gesture prototype was rated significantly better in terms of error recovery and in terms of error messages. This indicates that the type of error handling significantly influences user perception. Indeed, both systems follow different strategies: The PIN prototype supported corrective interaction and provided reset functionality. The gesture concept did not support undo operations. Consequently, every input error resulted in a failed attempt. In addition, the PIN prototype displayed an explicit pop-up window giving a binary error message (true/false). This window had to be explicitly confirmed, before the system was ready for the next attempt. In contrast, the gesture concept did not display any dialogue window but wrong patterns were shortly highlighted before the system was ready for the next attempt.

This indicates that, in contrast to focus on error prevention, the gesture concept prioritizes fast error recovery. We claim that such quick error recovery is especially useful in the context of mobile devices where input errors are likely but serious consequences cannot be assumed. Indeed, the user feedback indicates that quick recovery from errors is more important than error avoidance. In addition, the analysis of PIN-entry times revealed that error prevention is a time-consuming task. This trade-off between error prevention and speedy error recovery questions the importance of high success rates as a main usability factor for mobile authentication methods. Furthermore, we found that the feedback of the gesture-based concept was probably more comprehensible. As the prototype displayed the wrong pattern after a failed attempt, users could easily understand the source of error. In contrast, the PIN concept masked digits through dots and provided only binary feedback. This probably made errors harder to trace. Nevertheless, it should not be ignored that such design decisions can have implications for other aspects of the system (e.g., observability).

We conclude that the success rate is an important usability factor. However, effectiveness should not only be assessed based on the number of errors as some users deliberately fail to quickly start over. In general, we found that fast error recovery and comprehensible error messages are exceptionally important in the mobile context. This once again confirms that designers of user authentication mechanisms need to consider the context of use, especially when designing the process of handling errors.

Frequent Authentication Facilitates the Recall

Both concepts were easy to learn and both password types were easy to remember. Independently from the used concept, most participants learned their secret within three inputs. Consequently, most errors were based on inaccurate interaction and participants hardly had recall problems. The good memorability was finally confirmed by the spontaneous recall test. User feedback indicated that motor memory and visual aspects are very important for the memorization of the assigned tokens. While this may seem obvious when gestures are used, this was also the case for the PIN group.

The results suggest that the support of motor memory and visual memory are important design goals. However, we found no evidence for the assumption that the gesture-based approach is particularly easy to recall. For both concepts, we found that 21 repetitions are enough to achieve good learning effects. Since Section 3.2 indicated that users actually authenticate up to 50 times a day, we assume that, compared to other use cases where authentication takes place less often (e.g., ATM), memorability is a minor problem in the context of mobile devices. While this indicates that good performance should be prioritized, security tokens should still be designed in a way that makes them easy to remember.

The Benefit of Non-Restricted Gestures is Questionable

The theoretical security benefit of allowing a non-restricted gesture space is indisputable. Considering only the tested length of six cells, the number of unrestricted gestures already exceeds the number of Android-conform patterns by the factor of eleven. Considering gestures of all lengths makes the difference even larger. The quantitative analysis revealed that complex gestures were comparably easy to use and comparably memorable. However, we also found that participants who used complex gestures needed more time to authenticate. In addition, Chapter 2 already indicated that most users tend to select rather simple gestures. Hence, it is questionable if more complex gestures would actually be used in the wild. At the same time, allowing a non-restricted gesture set introduced input errors which would not have been possible with the (restricted) Android system.

We conclude that allowing non-restricted gestures increases the theoretical security without necessarily increasing practical security. At the same time, the results indicated that the theoretical security benefit is counterbalanced by practical usability drawbacks. We claim that usability should be prioritized to theoretical security benefits. However, slight usability drawbacks may be justified if they allow to significantly improve practical security.

3.3.4 Limitations

We collected usability data of PIN users and gesture users over the course of 21 days. The study was performed outside of the lab and participants were free to authenticate at any time of the day. However, we had to control various real-life factors to allow a valid comparison of both concepts. For example, we limited the task frequency to one authentication per day. We would like to note that most users are likely to authenticate more frequently in practice. However, we opted for low effort to minimize the dropout rates. Furthermore, we assigned secrets and controlled their length and complexity. We assume that most users would probably not opt for using a complex gesture, if self-selection is allowed.

The application itself was delivered as standalone tool and not integrated in the device's lock screen. We argue that the restrictive study design was necessary to increase the internal validity of the study and allowed a solid quantitative comparison. At the same time, the procedure clearly decreased the ecological validity of the study. We argue that the results gathered in this study are complementary to the insights gained from Section 3.2. Therefore, combining the results from both studies helps to get the full picture of both the usability and the utilization of current unlock methods.

Despite these limitations which were deliberately accepted to allow a sound comparison of both concepts, we have to point out further potential limitations which could not be avoided. First of all, we did not test interference effects. Even though the participants did not use their own credentials, it is very likely that most participants were trained PIN users and actively used various PINs on a daily basis (e.g., ATM). On the contrary, we cannot assume that all participants had gained previous experiences with gesture-based concepts. These differences may have influenced both the performance and the perception of the tested concepts. Finally, the pattern prototype allowed a wider range of gestures than Google's unlock concept. Consequentially, the results are representative for grid-based gestures in general but might not be generalizable to the specific Android unlock patterns.

3.3.5 Summary

In this Section, we presented the results of a 21-days field study which compared PIN and gesture-based unlock mechanisms. The controlled study design allowed to collect quantitative performance data and qualitative feedback. In addition, a spontaneous recall test gathered first insights into the memorability of both systems. Both concepts were compared in terms of *efficiency*, *effectiveness*, *perception*, *likeability* and *memorability*.

The results indicate that PINs outperform gestures when looking at the measured performance data. PINs were significantly faster and less error-prone. At the same time, gestures were perceived equally fast and easy to use. We found that one main difference between both concepts was the mechanism of error handling. While the PIN prototype provided undo functionality to correct failed input, the gesture-based authentication system did not support such operations. As a consequence, every wrong gesture was automatically logged as failed attempt. Even if this error handling method is likely to produce more logged input errors, we found that users did not desire undo operations. On the contrary, the gesture-based concept was rated significantly better in terms of error recovery. This indicates that, considering the mobile device context, quick error recovery is more important than error prevention. The results indicated that undo operations are often not used, even if failed attempts could be avoided. The recall test revealed similar memorability features for both concepts and indicated that recall problems are not a major issue if authentication systems are used on a daily base. The good recall rates were finally confirmed by the qualitative error analysis. The novel taxonomy revealed that most logged errors emerged from inaccurate input.

Overall, the study confirmed that gesture-based authentication methods are a usable and memorable alternative to traditional PIN-based authentication methods. Besides providing good usability, we found that most people were in favor of the pattern system and thought that gesture-based authentication feels good. This is a very promising finding as it increases the chance that gesture-based authentication systems will be adopted in the long run.

3.4 Observation Resistance of Grid-based Gestures

Section 3.2 presented important general requirements for usable unlock methods and Section 3.3 indicated that gesture-based authentication is a promising alternative to PIN. In this Section, we focus on the practical security of unlock gestures. While Chapter 2 already illustrated that the design of novel observation-resistant authentication methods has attracted much attention in the research community, the observability of currently used methods was not yet investigated systematically. Although it is often assumed that Android unlock gestures are prone to observation attacks, this vulnerability of grid-based gestures has not yet been quantified. This Section fills the gap in related work by providing answers to the following main *research questions*.

RQ1 How vulnerable are grid-based unlock gestures to observation attacks?

RQ2 Does gesture composition influence the vulnerability to observation attacks?

RQ3 Does gesture visualization influence the vulnerability to observation attacks?

To shed light on the research questions, we conducted an online study which simulated the gesture-based unlock process. Overall, 298 participants attacked 5960 unlock gestures of various length and complexity. Each attack consisted of an observation task and a drawing task. After each attack, we collected qualitative feedback concerning the difficulty of the performed tasks. The approach allowed to assess the relative impact of individual gesture characteristics such as length, knight moves, overlaps and visual appearance.

This Section presents the first systematic evaluation of the observability of grid-based unlock patterns and provides ground truth for their shoulder surfing vulnerability. The results indicate that unlock gestures are indeed prone to shoulder surfing attacks (*RQ1*). However, pattern composition has a significant impact on security (*RQ2*). While all tested parameters had significant influence, we found that pattern length is the most important factor. Furthermore, the results show that visualizing gesture strokes significantly downgrades security (*RQ3*). In addition to the quantitative analysis, we present a model that predicts the observability of a given unlock gesture. The contribution of this Section is twofold: Firstly, researchers can use the prediction model to further investigate the found trade-off and to develop feasible solutions. Secondly, gesture users can directly benefit by optimizing their used secrets according to the presented results.

This Section is partly based on von Zezschwitz, E., De Luca, A., Janssen, P., & Hussmann, H. (2014). Easy to Draw, but Hard to Trace?: On the Observability of Grid-based (Un)lock Patterns. In Proceedings of CHI'15. [265]. Please refer to the beginning of this thesis for a detailed statement of collaboration.

3.4.1 Research Context and Motivation

Chapter 2 revealed that the design of observation-resistant authentication concepts represents an active research field. To confirm the theoretically assumed observation resistance, most of such systems are tested in laboratory experiments where shoulder surfing attacks are simulated (e.g., [6, 27, 151]). Wiese and Roth [279] reviewed various shoulder surfing evaluations and found that such security studies are rarely performed systematically. The authors claim that different study designs and different procedures make the results hard to compare. For example, the input is sometimes performed by the experimenter and observation attacks are simulated by participants (e.g., [214]). In other studies, usability tests are filmed and the videos are used for a later expert analysis (e.g., [72]). Furthermore, the literature review in Section 2.3 indicated that most evaluations focus on the interaction with novel authentication systems and often forgo the comparison with baseline conditions (e.g., PIN, unlock gestures). Likewise, important factors like password length are often ignored. As a consequence, even though the vulnerability of current unlock methods is generally assumed, it was not yet examined systematically.

Zakaria et al. [292] analyzed the observability of the Draw-a-Secret scheme. The authors tested three symmetric shapes with the length of three, five and seven strokes. The user study followed a between groups design where each concept was tested by 17 participants. Participants observed the experimenter and reproduced the observed shapes using pen and paper. The analysis revealed two important findings: Firstly, password length played an important role as 100% of the 3-stroke shapes but only 52% of the 7-stroke shapes were successfully reproduced. Secondly, visual countermeasures like disappearing strokes helped to increase observation resistance. Schaub et al. [214] tested a mobile version of Pass-Go. They included short shapes (length = 2) and long shapes (length = 6) and found out that both lengths were successfully shoulder surfed in 70% of all observations. The input was performed by the experimenter and the observation was done by participants. Gugenheimer et al. [112] tested a novel concept called ColorSnakes. While the concept also provides more secure modes, the baseline condition is most relevant for this Section. The interaction with a direct path was comparable to the interaction with Android unlock patterns and can be used as a first indicator for the vulnerability of such gestures. Shoulder surfing attacks were simulated by three experimenters who reviewed the video footage of 24 participants. The authors simulated cognitive attacks (one-time observations) and video attacks. Depending on the grid size, the success rate was 96-100% when video attacks were allowed and 58-75% when single observations were simulated.

van Eekelen et al. [258] conducted an online survey to investigate the benefits of a newly developed authentication mechanism (PicassoPass) and included Android unlock pattern and PIN as a baseline. To simulate observation attacks, participants watched videos and selected one of six predefined answers. While in the gesture group, 13 of 17 (76.5%) attacks were successful, the success rate of PIN was even higher (94.4%). Unfortunately, the authors did not provide any information about the composition of the used gestures. Finally, the most relevant work was performed by Song et al. [232]. As part of the development process of a proactive password checker, a set of six unlock gestures was assigned to 101 participants. The set contained gestures of different strengths: Two gestures were rated “weak”, two gestures were considered “medium” and two gestures were rated “strong”. The strength was derived from equally weighting length, number of non-repeated segments and number of intersections. The unlock was performed by the experimenter and observed by the participants. After the observation, participants were asked to reproduce the observed gesture. 432 of 606 (71.29%) inputs were successfully reproduced. The analysis indicated a strong correlation between pattern complexity and observation resistance. Compromised gestures were significantly shorter, had significantly less intersections and less non-repeated segments. However, predefining the tested gestures limited the generalizability and thus the relative weights of composition aspects could not be analyzed in detail.

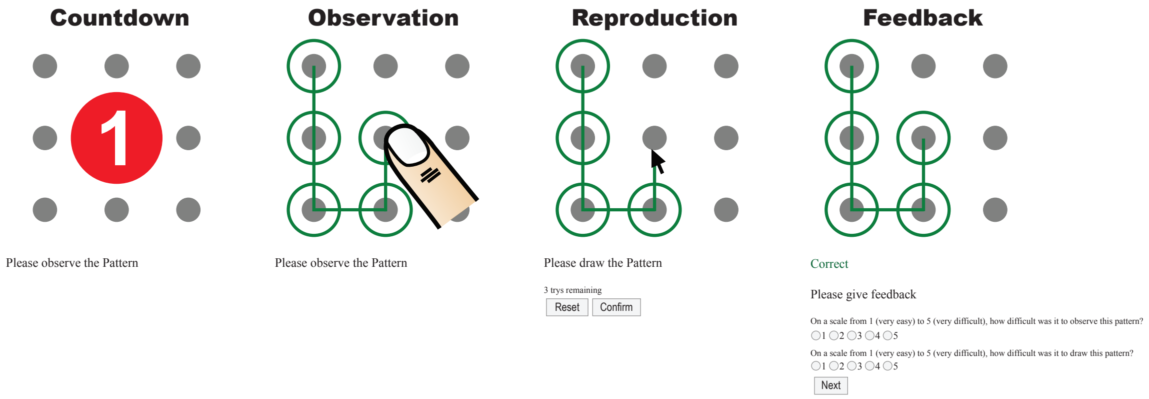
The literature review indicates that unlock gestures are vulnerable to observation attacks as most researchers report success rates around 70%. In addition, the results indicate that pattern composition and pattern length influence observation resistance. However, the relative weights of these influencing factors are still not well understood. In this Section, we fill the gap in related work by systematically investigating the observability of Android unlock patterns, the most popular gesture-based authentication system. We opted for an online study to maximize both the range of attackers and the diversity of the attacked gestures. Our approach allowed to weigh the most relevant composition aspects as well as the impact of stroke visibility. In summary, this Section presents the first large-scale analysis of shoulder surfing attacks on gesture-based authentication systems.

3.4.2 Online Study

The primary goal of the user study was to test a wide range of differently composed gestures. At the same time, we needed to guarantee that attack conditions were comparable across different participants. To maximize the range of tested gestures, we finally opted for machine generated patterns. We built a web-based study software which generated such patterns and then simulated the unlock behavior of humans. Participants observed the computer animated input and reproduced the observed patterns using their own devices.

Threat Model

Section 3.2 revealed that casual observations are most likely in the context of mobile authentication and therefore most relevant in the user’s daily life. We considered this finding in our threat model and assumed a casual observation in a semi-public space (e.g., public transport). The attacker has exactly one opportunity to observe the gesture input as there is no technical equipment involved. The attacker has perfect sight on the user’s device and the whole authentication process can be observed. In addition, the attacker has no previous knowledge about the specific characteristics (e.g., length) of the performed gesture. After performing the shoulder surfing attack, the attacker gets into physical possession of the device and tries to reproduce the observed pattern.



Please observe the Pattern

Please observe the Pattern

Please draw the Pattern

3 trys remaining

Reset Confirm

Correct

Please give feedback

On a scale from 1 (very easy) to 5 (very difficult), how difficult was it to observe this pattern?

○1 ○2 ○3 ○4 ○5

On a scale from 1 (very easy) to 5 (very difficult), how difficult was it to draw this pattern?

○1 ○2 ○3 ○4 ○5

Next

Figure 3.16: The study interface consisted of a 3 $*$ 3 matrix. After a three seconds countdown, the gesture was displayed. Next, participants reproduced the gesture. After the attempt was submitted, participants rated the task using five-point scales.

Study Software

The front end of the study software was based on JavaScript; the back end was created using PhP and MySQL. The front end, which simulated a perfect view from above, is illustrated in Figure 3.16. Each observation started with a three seconds countdown. Afterwards, the animated authentication was displayed. Next, participants had three attempts to successfully reproduce the observed pattern before feedback was collected.

The 3 $*$ 3 matrix was displayed with an edge-size of 500px. As seen in Figure 3.16, input was additionally indicated using a virtual finger. This finger was the only visual indicator whenever stroke visualization was deactivated. The animation ran with a fixed speed of 500ms per single stroke. The speed was derived from Section 3.3 which indicated an average real-world speed of 2.7 seconds for Android-conform five-stroke gestures. Due to the larger distance between single cells, overlaps and knight moves took more time. User input matched the standards known from common graphic editors as participants started drawing by pressing the left mouse button and finished by releasing it. User interaction was logged and sent to the back end.

Method

The study was conducted following a repeated measures design. Each session consisted of an introduction, a training task and 20 observations. The introduction page explained all important aspects of the user study. We gave details on the study goal, explained observation attacks as well as the procedure and the interaction. After the introduction, a training task was performed. The training task consisted of three gestures of different complexity. Finally, the user study started and each participant was asked to complete 20 observation attacks.

Independent Variables
Gestures were randomly generated based on the standard rules for Android devices. We defined the following independent variables:

Line visibility [true/false] – specifics if a drawn pattern (strokes) is visualized or not. The factor was alternated between the attacks and therefore "true" was assigned to exactly 50% of the gestures and vice versa. As a result, we were able to test similar gestures (with similar complexity) with both, visible and invisible lines.

Length [4-9] – specifies the number of activated cells. According to the standard rules for Android devices, patterns comprised a minimum number of four cells and a maximum of nine cells. The number of cells was randomly assigned.

Knight Move [0-4] – specifies the connection of two distant cells which are not directly neighbored (e.g.,). According to the standard rules of Android devices, only straight lines are allowed (cells on a straight line cannot be skipped). Knight moves were assigned with a probability of 20%, whenever possible.

Overlap [0-3] – specifies crossing over an already activated cell by connecting to a distant cell (e.g.,). Overlaps were assigned with a probability of 20%, whenever possible.

Intersection [0-7] – specifies strokes which cross already drawn strokes (e.g.,). Intersections were randomly assigned.

Dependent Variables
The main dependent variable was shoulder surfing success. It was measured in two ways:

Binary Success [true/false] – specifies if a gesture was successfully observed. If binary success equals "*true*", the attacker would have been able to authenticate.

Percentaged Success [0-100] – specifies the portion of successfully observed cells. The percentaged success rate is given by the sum of correctly observed cells divided by the length of the observed gesture (entirety of correct cells). A cell is correctly observed when both the spatial position in the matrix and the temporal position within the gesture match the values of the expected cell. Higher values indicate a higher chance for the attacker to authenticate. A value of 100% equals "binary success = true".

Procedure
Each participant performed the following steps. The whole procedure took 13 minutes (SD = 5) on average.

Introduction Reading the introduction page.

Training Observing and reproducing three gestures of different complexity. All participants tested the same three gestures.

Attack Observing and reproducing 20 gestures. Each gesture was observed exactly once. For reproduction attempts, correction was possible. A maximum of three attempts was granted. After each attack, feedback was collected using five-point scales.

Questionnaire After 20 attacks, participants were forwarded to a questionnaire. We collected demographical data and assessed relevant previous experiences (e.g., shoulder surfing). Furthermore, the questionnaire investigated if additional equipment was used during the study (e.g., pen and paper).

Raffle All participants had the chance to win one of two eBook readers. The chance of winning was positively correlated with the number of successful observations. We assumed that this aspect would additionally motivate participants.

Multiple participations were forbidden. If the use of additional equipment was reported, the respective participant was excluded from the analysis. However, such participants still had the chance to win one of the eBook readers.

Participants

Participants were recruited using a university-wide mailing list. As two of 300 participants reported to have used additional equipment, the final data set was based on 298 participants. The proportion of male (51%) and female (49%) participants was well-balanced. Participants indicated an average age of 32 years (14-73, SD = 13). While most (59%) used Android smartphones on a daily base, about one third (29%) reported the use of other smartphones (e.g., Apple iPhone). In addition, 12% of the sample reported that no smartphone was used on a daily base.

When asked about the current lock screen configuration, 30% reported to use PIN, 28% used the Android gesture unlock and 11% indicated other methods (e.g., fingerprint). Further 31% did not use a secure lock screen (e.g., slide-to-unlock). While 15 (5.0%) participants had already been a victim of a shoulder surfing attack, 44 (14.8%) reported that they already had observed code entries of others.

Results

Overall, we collected 5960 (20 observations $*$ 298 participants) samples. However, we had to remove 61 patterns as their characteristics exceeded the defined maximum of the independent variables[11]. That is, these patterns comprised more than four knight moves, more than three overlaps or more than seven intersections. Including such extreme values would have skewed the results as the number of observations was insufficient to allow a valid analysis.

Table 3.3 shows the main statistics of the final pattern set. The data set comprised very simple patterns which are very likely to be used by humans (e.g., 12% of the patterns with

[11] The removed patterns comprised extreme values which had been generated too infrequently to allow statistical analyses.

Factor	Mean (SD)	Median	Distribution
Line	-	-	false [50.0%], true [50.0%]
Length	6.36 (1.72)	6,00	4 [19.5%], 5 [17.4%], 6 [16.5%], 7 [16.4%], 8 [15.4%], 9 [14.9%]
Knight move	0.92 (0.93)	1,00	0 [39.2%], 1 [37.4%], 2 [17.0%], 3 [5.1%], 4 [1.2%]
Overlap	0.38 (0.65)	0,00	0 [69.9%], 1 [23.5%], 2 [5.3%], 3 [1.4%]
Intersection	1.06 (1.39)	1,00	0 [48.4%], 1 [24.4%], 2 [12.3%], 3 [7.8%], 4 [3.9%], 5 [1.8%], 6 [1.0%], 7 [0.3%]

Table 3.3: Main statistics of the independent variables of the final data set.

visualized strokes did not comprise a special move) but also more complex patterns which are unlikely to be used in the wild. A first analysis revealed that the outcome of the first attempt of reproducing a pattern was a very good predictor for the overall success. That is, in most cases (94%) where participants failed to reproduce the pattern in the first attempt, the attack remained ineffective even after three attempts. As a consequence, we focus on the analysis of participants' first guesses.

Binary Feature Weights

To assess the relative feature weights on a binary basis, we define *success* as *false* (coded as 0) and *true* (coded as 1). Overall, 3565 (51.7%) patterns were successfully shoulder surfed, 57.9% of them had visible lines. In contrast, 62.2% of the resistant gestures had line visualization disabled. Individual independent t-tests for each pattern feature and *success* reveal that shoulder surfing resistant gestures comprised significantly more cells (M = 7.4, SD = 1.4; $t_{5897} = 44.5$) than exposed ones (M = 5.7, SD = 1.5). In addition, resistant gestures comprised significantly more *knight moves* (M = 1.3, SD = 1.0) than exposed ones (M = 0.7, SD = 0.8), more *overlaps* (M = 0.6, SD = 0.8) than exposed ones (M = 0.3, SD = 0.5) and more *intersections* (M = 1.7, SD = 1.6) than exposed ones (M = 0.6, SD = 1.0), all $p < .001$.

To determine the relative weights of the different factors, we conducted a binary logistic regression analysis. The main results are depicted in Table 3.4 (left). The analysis reveals that all tested factors have a significant individual impact on the observation resistance of a given pattern (all $p < .001$). The resulting prediction model is able to correctly estimate 75.8% of the binary outcome of an observation attack ($\chi^2(5) = 20089.9$, $p < .001$, $R^2_{Nagelkerke} = 0.404$). The odds ratio (see Table 3.4) reveals that switching off line visualization reduces the chance of a successful observation attack by 67%. Furthermore, each additional cell reduces the risk that the pattern is observed correctly by 45%. The use of a knight move reduces the risk by 32% and each overlap lowers the chance of a successful attack by 20%. Finally, each intersection reduces the risk of shoulder surfing by 12%.

Figure 3.17 illustrates the measured binary success rates for each composition factor. While increased complexity generally improves observation resistance, switching off line visualization has a significant impact at any level.

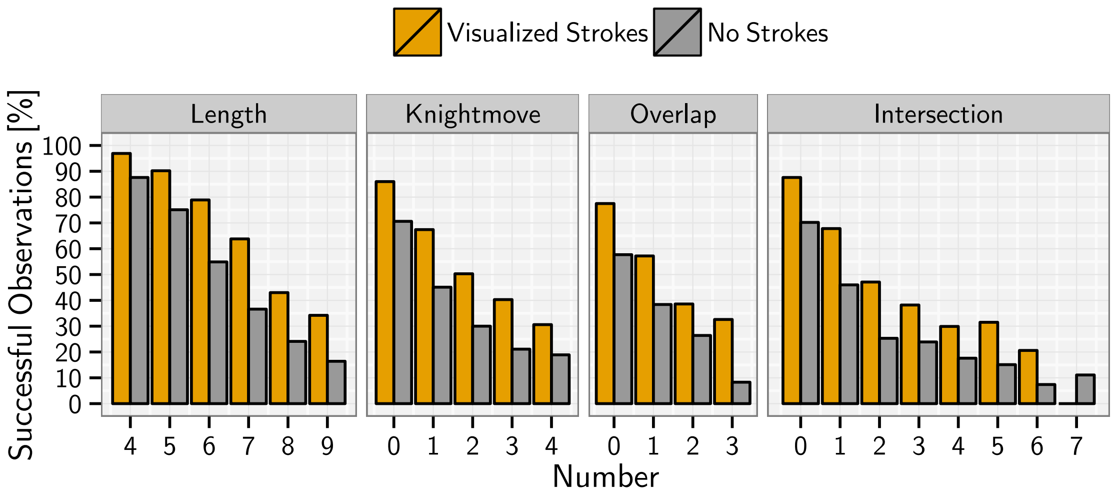


Figure 3.17: *Binary success* rates subdivided by composition factors and line visualization. While each factor has a significant impact, length and line visualization are most important.

	Binary Logistic Model		Linear Regression Model		
	B (SE)	Odds Ratio (95% CI)	B (SE)	β	VIF
Line	-1.12 (.07)	0.33* (.29, .37)	14.42 (.69)	.23*	1.00
Length	-0.60 (.026)	0.55* (.52, .58)	-5.27 (.27)	-.29*	1.78
Knight move	-0.38 (.05)	0.68* (.62, .75)	-3.99 (.53)	-.12*	2.02
Overlap	-0.22 (.05)	0.80* (.72, .89)	-3.20 (.60)	-.07*	1.27
Intersection	-0.13 (.04)	0.88* (.82, .95)	-2.05 (.39)	-.09*	2.52
Constant	5.52 (.17)	-	97.59 (1.85)	-	-

Table 3.4: Left: B-values and odds ratio of the logistic regression model predicting *binary success*; Right: B-values, standardized betas and variance inflation factor of the linear regression model predicting the *percentaged success* rate. Line visibility was coded: 0 = false, 1 = true. All tested features have a significant individual impact on observation resistance ($*p < .001$).

Success Rate Prediction

Despite the *binary success*, we defined the *percentaged success* rate as the portion of correctly reproduced cells. Values can range between zero (no correct cell) and 100 (all cells are correct). Analyzing the portion of correct cells helps to specify the observation risk of a pattern even if it was not correctly reproduced. Therefore, *percentaged success* is an important measure to assess the difficulty of the observation.

Overall, participants were able to correctly reproduce 78.8% (SD = 30.9) of the observed cells (patterns). Analyzing the group of observation resistant patterns reveals that participants were overall able to observe 46.4% (SD = 28.1) of the input, even if the overall outcome was wrong. In addition, we found that participants were able to correctly reproduce 71.4% (SD = 34.4) of the cells, when line visualization was deactivated. Switching on line visualization increased the portion of success to 86.2% (SD = 24.9).

We performed a simple multiple regression analysis to specify the relative weights of each factor and to predict the observation risk for a given pattern. The data met the assumption of independent errors (Durbin-Watson value = 2.035). Furthermore, preliminary analyses indicated no multicollinearity and the histogram as well as the P-P plot of standardized residuals indicated that errors were approximately normally distributed. The analysis resulted in a highly significant regression equation ($R^2 = 0.263, R^2_{Adjusted} = 0.263, F_{(5,5893)} = 421.32$, $p < .001$). The details of the model are depicted in Table 3.4 (right). The linear regression model confirms the findings of the binary analysis. All factors are significant individual predictors for observation risk ($p < .001$). Looking at the standardized β-values indicates that the number of cells has the biggest impact ($\beta = .29$), followed by line visibility ($\beta = .23$), knight moves ($\beta = .12$), intersections ($\beta = .09$) and overlaps ($\beta = .07$).

Therefore, the observation resistance of a given pattern can be assessed by the following *equation*:

$$\begin{aligned}
&ObservationRisk = && \mid \textit{portion of correctly observed cells}\\
&\quad 97.59 && \mid \textit{constant term of the prediction model}\\
&\quad +14.42 * X_{line} && \mid X = \textit{1 if visualization on, else } X = 0\\
&\quad -5.27 * X_{cells} && \mid X = \textit{ number of cells}\\
&\quad -3.99 * X_{knight} && \mid X = \textit{ number knight moves}\\
&\quad -3.20 * X_{over} && \mid X = \textit{ number of overlaps}\\
&\quad -2.05 * X_{inter} && \mid X = \textit{ number of intersections}
\end{aligned}$$

Further analyses revealed no significant impact for personal attributes like gender or the daily use of unlock patterns.

User Feedback

After each observation, participants rated the ease of observation and the ease of input based on two five-point scales ranging from 1 ("very easy") to 5 ("very hard").

A Spearman's rank-order correlation indicates a strong positive correlation of the answers to both questions, $r_s(5897) = 0.96, p < .001$. Observation resistant patterns were rated both harder to observe (Mdn = 5) and harder to reproduce (Mdn = 5). At the same time, both tasks were rated easy (Mdn = 2) for patterns which were successfully attacked.

Overall, user feedback confirms that increasing complexity makes both tasks more difficult. While short patterns with four cells were rated "very easy" (Mdn = 1) to draw and "very easy" (Mdn = 1) to observe, both tasks were rated "medium" (Mdn = 3) for patterns with six cells and "very hard" (Mdn = 5) for gestures with nine cells. The rating of patterns which comprised knight moves, intersection and overlaps showed the same trend. Finally, both tasks were rated "medium" (Mdn = 3) when gestures were visualized and "hard" (Mdn = 4) when strokes were invisible.

3.4.3 Discussion and Implications

We now summarize the results gained from the online study and discuss relevant implications for both users and researchers.

Unlock Gestures are Prone to Observation Attacks

After observing the input once, participants were able to reproduce 51.7% of all tested patterns. Even if the correct pattern was not reproduced, users were still able to recognize almost half of the cells correctly. In practice, attackers could substitute missing cells with additional information. For example, parts of the gesture could be gained from smudge left on the screen or derived from known user preferences. Therefore, the results confirm that Android unlock gestures are vulnerable to observation attacks, even in one-time observations. In addition, we assume that this is especially the case if multiple observations are possible or video attacks are performed.

Gesture Composition and System Configuration Are Important Factors

Even if unlock gestures are generally easy to attack, we found that all tested factors have a significant impact on observation resistance. The number of cells and line visibility are the most important aspects. When users switch off line visualization, the chance of a successful observation attack is immediately reduced by 67%. Every additional cell lowers the risk of an observation attack by 45%. In addition, including "special moves" like overlaps or intersections can significantly reduce the shoulder surfing risk. However, even if such long, complex and invisible patterns are harder to observe, they are hardly a practical solution as the use of such gestures increases error rates and input times. Therefore, we claim that novel authentication mechanisms need to be found which provide a better trade-off between observation resistance and usability.

Real-world Patterns are Particularly Easy to Observe

Prior work already indicated that users tend to select simple and short gestures which are fast and easy to use [10, 254]. This is confirmed by the fact that participants from our study rated more complex patterns harder to draw. In addition, as line visibility is usually activated per default[12], it can be assumed that most users draw visualized lines in practice. Therefore, we assume that long, complex and invisible gestures are hardly used in the wild. When considering such gestures which are likely to be used in the wild, we find that 93% of such patterns with visualized lines and without any "special moves" were successfully reproduced in our experiment. We therefore conclude that real-world gestures are particularly easy to observe.

[12] `http://www.tech-recipes.com/rx/32304/android-how-to-hide-patterns-unlock-phone/` – last accessed: 2015/12/29

Some Gestures Might be Easy to Enter but Hard to Trace

Even if patterns which are harder to observe are often harder to enter, we assume that some "more secure" patterns can be still fast to use. While knight moves demand more accurate interaction, overlaps and intersections do generally not demand additional input effort. As the reproduction of gesture-based authentication is mainly based on motor memory, we furthermore assume that most users would quickly get used to invisible patterns. Our prediction model can be used to further investigate the interplay of usability and observability. For example, it can be integrated in a proactive pattern meter which visualizes the estimated shoulder surfing risk for a given pattern. Such systems could be displayed during pattern selection and help users to avoid high risk gestures. At the same time, researchers could use the system to systematically investigate patterns which are harder to trace, but still easy to perform.

3.4.4 Limitations

The web-based study software, which controlled the pattern generation and simulated the user input, allowed that a large number of patterns was observed by many different attackers. Thus, the specific study design eliminated the effects of different observers and different patterns. At the same time, we had to exclude some real-world factors. The input was performed at a constant speed which was derived from field evaluations. This input speed would vary in the real world. In addition, the study simulated a best case scenario for attackers as they had perfect view from above. In reality, the angle of view would change and the view could be disturbed by reflections or occlusions. Finally, the simulation was limited to right-handed input. Thus, it is possible that the actual weight of individual factors will change depending on the context and the performance of the user. However, we assume that the relation of single features will stay the same. At the same time, we assume that most real-life factors (e.g., faster input) are likely to reduce reproduction rates. Finally, the prediction model is only applicable to one-time observations. Nevertheless, the results do not indicate that any of the tested features provides significant protection from more advanced attacks.

Despite the specific limitations of the study procedure, online studies always comprise a certain lack of control. For example, we could not control the study environment as participants performed the observations using their own devices. Furthermore, we could not control all participants' characteristics. For example, it is possible that long-sighted users performed the task without glasses. Finally, we cannot guarantee that all participants reported the use of additional equipment. However, as we recruited a very large number of participants, we are confident that such external factors did not significantly influence the results.

3.4.5 Summary

In this Section, we presented a systematic evaluation of the observation resistance of unlock gestures and provided ground truth for their real world vulnerability. However, the analysis of 5960 observation attacks did also show that each composition factor plays a significant role. That is, the results indicated that using long invisible gestures would significantly reduce the observability of the system. Finally, we presented a regression model that can be used to predict the observability of a given pattern.

This Section provided valuable insights for current systems and for future developments. First, we contributed to the understanding of the observability of unlock gestures and presented solutions that can be immediately applied to current systems. Secondly, the results confirmed that currently used methods are often very easy to observe. While the use of long, invisible and more complex gestures may somewhat reduce observability, the results gained from Section 3.2 and Section 3.3 do not indicate that such strategies will be widely accepted. Therefore, we conclude that novel gesture-based authentication mechanisms need to be developed. This novel class of gesture-based authentication mechanisms should combine effective observation resistance with efficient and satisfying input methods.

3.5 Practical Password Space of Grid-based Gestures

Section 3.2 and Section 3.3 showed that unlock gestures are particularly popular among users as they are fast and easy to use. However, Section 3.4 showed that such systems are prone to observation attacks. In addition to observability, Chapter 2 already indicated that guessability might be a problem as many users follow predictable selection strategies. Nevertheless, the actual strength of a given unlock gesture is still hard to define.

In this Section, we fill this gap of knowledge and present a novel metric that enables to quantify and to compare the similarity of grid-based unlock gestures. We apply our metric to a user-defined gesture set and provide answers to the following main *research questions*.

RQ1 How can we quantify the practical space of grid-based unlock gestures?

RQ2 How predictable are user-defined grid-based unlock gestures?

RQ3 Which are the most popular selection strategies and what are popular unlock gestures?

To measure the diversity of user-defined unlock patterns, we introduce a novel metric for grid-based patterns which is based on geometric similarity and the assumption that unpopular patterns are generally more resistant to guessing attacks [167,215]. The metric is inspired by Li and Vitányis's use of the Kolmogorov Similarity measure [163] and quantifies similarity through the computational efforts required to derive a given representation of a gesture A from another gesture B (*RQ1*).

We adopt the metric in a greedy clustering approach and measure the diversity of 506 user-defined unlock patterns. The approach reveals popular gesture groups, which are often selected by users. The gestures of such clusters show high similarity values in terms of shapes and complexity. We will illustrate that user-defined grid-based unlock gestures can be clustered to a small number of groups and are therefore very predictable (*RQ2*). We show that applying up to two simple transformations (e.g., rotation) already reduces the number of distinct gestures by about two thirds. Further analysis of popular patterns reveals that users are indeed in favor of short and simple shapes. For example, almost one third of all user-defined gestures could be derived from five central patterns (*RQ3*).

Overall, this Section confirms that user-defined grid-based unlock gestures are very much prone to dictionary attacks. Beyond that, the presented strength metric builds the basis for a comparable evaluation of future developments. For example, it will be used in Chapter 4 to evaluate and compare the effectiveness of proactive pattern checkers.

This Section has been partly included into von Zezschwitz, E., Eiband, M., Buschek, D., Oberhuber, S., De Luca, A., Alt, F. & Hussmann, H. On Quantifying the Effective Password Space of Grid-based Unlock Gestures. To appear in Proceedings of MUM'16 [267]. Please refer to the beginning of this thesis for a detailed statement of collaboration.

3.5.1 Research Context and Motivation

Chapter 2 provides an in-depth discussion of guessing attacks on gesture-based authentication mechanisms. Previous work already indicated that unlock gestures are easy to guess [254]. Uellenbeck et al. [254] reported that most patterns are drawn from left to right and that users prefer the upper left cell as a starting point. Andriotis et al. [11] confirmed that a biased selection behavior limits the practical password space of the authentication system. In addition, the predictability of Android patterns has recently been covered by the media[13].

Proactive security checking was already discussed as a possible solution [10, 229, 232, 241]. Such systems work analogously to alphanumeric password meters and calculate the strength of a given gesture based on specific composition aspects. Proposed features included length [10, 232, 241], direction changes [10] or knight moves [10]. However, the actual weight of these measures was not yet investigated and thus it remained unknown what exactly makes a grid-based gesture hard to guess. In addition, it was already shown that statistical composition measures do not reflect actual user choice [97, 145, 177].

As a consequence, previous work proposed simulated guessing attacks as an alternative approach to quantify guessability. The performance of such guessing algorithms depends on appropriate training data [273]. In the context of alphanumeric passwords, such data can be collected from exposed databases. However, as unlock gestures are usually stored on the owner's device, researchers have to rely on training sets which are often collected during simulated enrolments. We claim that the use of self-collected training sets questions the comparability between different projects.

In this Section, we provide a novel metric which does neither require training data nor relies on statistical assessments. As a consequence, the metric allows to compare the practical password space of arbitrary gesture sets, even if they were collected under different conditions. The metric is based on geometric similarity and the assumption that unpopular patterns are generally harder to guess [167, 215]. Therefore, two factors are particularly important: A) the similarity to known popular patterns and B) the similarity to other patterns in the same gesture set. The more a given gesture differs from A) and B), the harder it is to guess. Gesture sets which show less similarity represent a less predictable and thus more secure practical password space.

3.5.2 Similarity Metric

We compare patterns according to their geometric similarity. In this Section, we define the similarity metric and present the clustering approach.

[13] `https://nakedsecurity.sophos.com/2015/08/22/surprise-people-choose-predictable-android-lockscreen-patterns/` – last accessed: 2016/01/05.

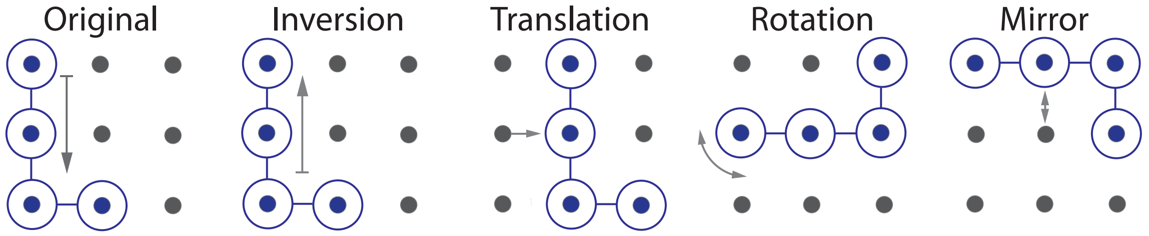


Figure 3.18: All five gestures are based on a simple L-shape. The similarity metric proposed in this Section assesses the distance between such gestures by analyzing the number of geometric transformations needed to convert one gesture into another.

Definition

We aim to determine the portion of similar shapes within a given set of user-defined unlock gestures. Therefore, we first define "similarity" inspired by Euclidean plane isometries [49]:

> *Two patterns A and B are n-similar for $n \in \mathbb{N}_0$ respecting a set of transformations T, if A can be transformed into B with exactly n geometrical or logical transformations from T. If A and B are equal, they are assigned a distance of* 0.

We define the following transformations T. Figure 3.18 illustrates examples for each transformation.

Inversion Traverse a shape in inverted order
Translation Translate a pattern by one cell in north, east, south or west direction
Rotation Rotate a gesture by 90, 180 or 270 degrees
Mirror Mirror a pattern on the x-axis or y-axis

Example If we choose $n = 1$, this means that patterns of the same group are "1"-similar. In other words, each gesture of the group could be transformed into a specific "central" gesture of the group by applying exactly one transformation T.

As proposed by Li and Vitányis [163], we quantify similarity through the computational efforts required to derive a given representation of a gesture A from another gesture B . Thus a higher number of operations translates to less similar gestures.

Gesture Grouping Approach

We limit the analysis to groups of congruent gestures (with equal length), since we consider shapes to be the most important property in the pattern creation process, and a change in length often alters the shape of a pattern. Each group contains all patterns n-similar to a "central" pattern within the group. Furthermore, each pattern is assigned to exactly one group. Therefore, choosing the central patterns in a way that minimizes the total number of groups presents an optimization problem.

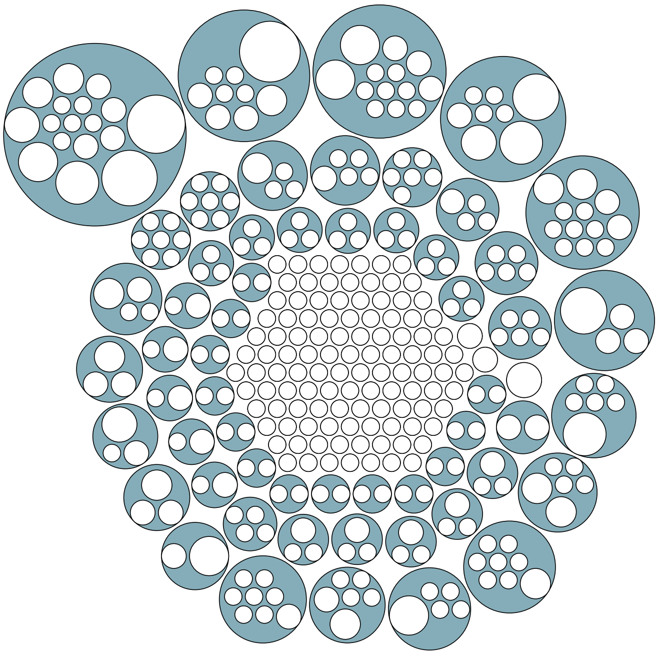


Figure 3.19: The results for $n \leq 2$. Each white circle is a pattern; its size grows with its total occurrence count in the database. Dark circles enclose groups of similar patterns.

Example Problem To motivate and explain the clustering approach, we start with the following exemplary problem:

$$A \xrightarrow{1} B \xrightarrow{1} C$$

with $n = 1$ for the pairs $\{(A,B),(B,C)\}$. In practice, we could imagine transformations of the popular "L"-shape. A would be an "L" aligned to the left ( ), B could be the same "L" translated right ( ), and C would be the same as B but rotated by 90 degrees ( ). This means that A has a distance of two to C, while B has a distance of one to both. Thus, the optimization problem becomes evident: If we choose $n = 1$ (i.e., the maximum distance within groups is 1) and start with A as a *central* pattern, we will get two groups $\{A,B\}$ and $\{C\}$. If we use B as the central pattern instead, we would only get a single group $\{A,B,C\}$ to cover all three shapes. Figure 3.18 illustrates another example: The distance of the *original* and the *inversion* as well as the *translation* (assuming the same direction) is $n = 1$, the *rotation* and the *original* differ by $n = 2$, the distance of the *mirrored* gesture and the *original* is $n = 3$.

Approach - Greedy Clustering Generating the optimal set of central patterns is equivalent to the Minimum Set Covering problem [60], and thus NP-hard. Therefore, finding the *optimal* solution for gesture sets of interesting size is not feasible as it would either take too much time or require too much computer power. Consequently, we opt to approximate the optimal solution by using a greedy algorithm [156] instead.

In each step, we add one ungrouped gesture to the set of central gestures. We choose the gesture which is n-similar to the largest number of yet ungrouped gestures. This procedure is repeated until each gesture is part of one group, possibly a group containing only one gesture (i.e., a unique one, which cannot be derived with n transformations from any other gesture).

As we are particularly interested in very similar gestures, we consider a maximum of n=2. Figure 3.19 visualizes[14] the results of the greedy clustering applied to an exemplary dataset which was collected under standard Android conditions. It shows that despite a number of unique patterns (in the center), the vast majority of patterns belongs to groups, meaning that they can be derived from each other within $n \leq 2$ simple transformations. The next Section presents the online study and discusses its results in detail.

3.5.3 Online Study

The similarity metric is applied to a set of 506 user-defined unlock patterns which were collected under Android standard conditions.

Threat Model

Even though the main purpose of this Section is to introduce a novel metric for pattern strength, we present a potential threat model. According to our threat model, the attacker is in possession of a device which is protected by a grid-based unlock gesture. The attacker has no previous knowledge about the used gesture. However, the attacker knows the most popular Android patterns. As a consequence, she performs a dictionary attack and starts with the most popular pattern. Whenever the attack fails, the attacker tries another gesture which results from simple geometric transformations. The smartphone allows up to 20 guesses within a feasible amount of time (i.e., a few minutes), before the device gets blocked for a longer period. It should be noted that the search space is further reduced, if the attacker is able to gain additional information about the characteristics of the used gesture (e.g., by smudge analysis or by observation).

Method

The goal of the study was to collect a huge set of user-defined gestures from users of various backgrounds. To quickly collect such patterns, we developed a web application for mobile devices and recruited participants using Amazon Mechanical Turk (MTurk). The study started with an introduction page that explained the topic of the evaluation and provided details of the procedure. After the task was accepted, a link to our web application was displayed. MTurk users were asked to open this URL on their smartphone. We utilized PHP Mobile Detect[15] to examine if participants actually used their mobile devices.

[14] Generated with `http://d3js.org` Library released under BSD license. Copyright 2015 Mike Bostock.

[15] PHP Mobile Detect released under MITLicense. (`http://mobiledetect.net`) – last accessed: 2015/11/15.

Both the interaction and the graphical appearance of the user interface resembled current implementations of the Android unlock system. The pattern creation process followed the standard Android enrollment procedure[16]. That is, the patterns had to be conforming to the standard Android rules and selected gestures had to be confirmed once. During the selection process, participants were allowed to reset their input and start again. After the gesture was confirmed and submitted, we displayed a short questionnaire which collected demographical data and gathered information on the technical background. After the successful completion of all steps, a secret code was provided. Participants had to enter this code in MTurk to confirm the completion of the study. The whole procedure took 102 seconds on average (SD = 53). Each participant was compensated with US$ 0.5.

Participants

We checked that all participants provided the correct confirmation code. In addition, we validated the given answers and excluded participants who did not fulfill the requirements. For example, we excluded participants who stated not to use mobile devices. Finally, 506 participants contributed to the data set. All participants indicated to be US citizens, 334 were male and 172 were female. The average age was 28 years (SD = 8; 18-67). 50.2% reported to use Android smartphones, 49.4% used iPhones and two participants used Blackberry devices. Overall, 38.7% indicated to use PIN on their smartphone, 37.5% used no secure lock screen, 17.7% used Android unlock gestures and 6.1% used other methods.

Results

Since each participant contributed one sample, the results are based on 506 unlock gestures.

Basic Statistics

The average pattern length was 5.0 points (SD = 1.4). The basic statistics confirmed the findings of related work [11,254] as the favored starting point was at the top left with 41.1%. Most patterns followed the western reading direction and 20.0% finished at the bottom right. Most gestures were based on simple strokes. Special moves were hardly used: Only 7.3% of the patterns included overlapping nodes and only 5.9% comprised knight moves.

Popular Gestures

The results of the similarity analysis confirm that users select their unlock patterns from a limited pool of simple shapes. The groups for $n \leq 2$ with pattern length four deserve special attention, as they include more than half of the patterns in the dataset. The largest observed group is formed around the L-shape (Table 3.5) and covers 17 different permutations.

Hence, attackers brute-forcing their way through all these 17 permutations of L will get a hit for 56 of the dataset's 506 patterns – that is 11.1%. The second largest group comprises N-forms and covers nine different patterns, whose occurrences account for almost 5% of our

[16] More details: `http://phandroid.com/2014/03/20/android-101-lock-screen/` – last accessed: 2015/09/17.

Rank	Top Gesture	# Permutations	# Occurrences	% Dataset
1	(1478)	17	56	11.1
2	(1596)	9	25	4.9
3	(5896)	13	24	4.7
4	(1256)	8	23	4.5
5	(1235)	12	18	3.6

Table 3.5: The five largest groups for $n \leq 2$. The table shows the most frequent (top) pattern of each group, the number of different patterns covered in the group, the accumulated absolute number of occurrences for all patterns in the group, and the covered ratio of the whole dataset.

dataset. The group ranked three includes and covers 13 different patterns with 24 occurrences in our data (4.7%). Overall, the five largest groups presented in Table 3.5 comprise 59 different patterns. Their occurrences account for roughly 29% of the dataset.

Considering gestures which comprise more than four cells, reveals that the largest group contains patterns of length seven. It is the sixth largest group in the whole data set. The group covers 13 -shapes in four different permutations. This means that over 40% of all length-seven-patterns can be derived from this single form. Most (8.4%) of the length-five-patterns are based on -forms, 23.3% of the length-nine-patterns form -shapes. Patterns with six or eight cells show most diversity.

Gesture Similarity

Figure 3.20 and Table 3.6 summarize the results: The total number of distinct gestures shrinks from 506 to 350 when removing duplicates (i.e., $n = 0$). For a similarity distance of 1, the total number of individual groups is 213. Hence, considering those patterns as duplicates that differ only by a single transformation already reduces the number of unique patterns by about 57%. If we set $n \leq 2$, only 169 groups are left, as determined by the greedy grouping algorithm. That is, considering very similar gestures ($n \leq 2$) as duplicates would shrink the practical gesture space to a third of its size.

Gestures with length ≥ 5 show less similarity than gestures with only four cells. However, the results in Table 3.6 imply that the ratio between the total gesture count and number of similarity groups does not shrink linearly with pattern length. Instead, it reaches a minimum for length six, where the number of diverse patterns is reduced by just 33%. Length eight is close second with a reduction by 40%, and the unique pattern count for length nine – where one might expect the biggest reduction – is reduced by 43%, indicating similar selection strategies.

If we focus on the more than 50% of participants who used patterns with a length of four, we find exceptionally high similarity values: As indicated by Table 3.6, the grouping algorithm clusters 262 gestures to 44 similarity groups with $n \leq 2$ and thus reduces the practical gesture space to 17% of its size.

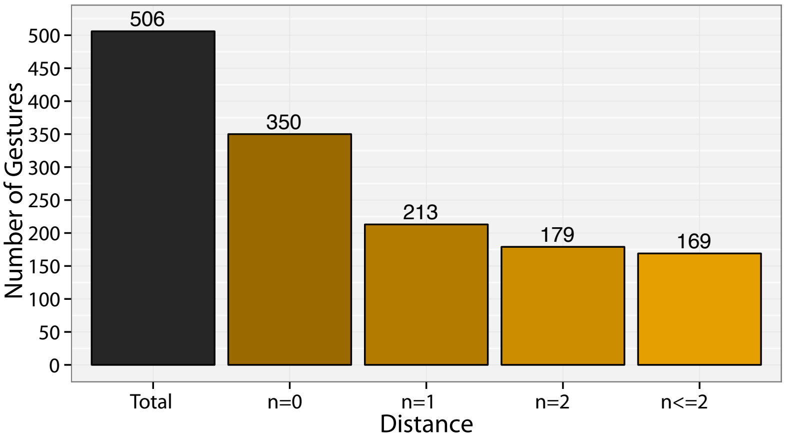


Figure 3.20: Total number of unique gestures (groups) for $n \in \{0,1,2\}$. Considering a distance of $n \leq 2$ as duplicates reduces the gesture space by 66%.

Length	Total		n = 0		n = 1		n = 2		n≤2	
any	506	*100%*	350	*69.2%*	213	*42.1%*	179	*35.4%*	169	*33.4%*
4	262	*51.8%*	156	*30.8%*	68	*13.4%*	50	*9.9%*	44	*8.7%*
5	119	*23.5%*	94	*18.6%*	64	*12.6%*	56	*11.1%*	52	*10.3%*
6	48	*9.5%*	43	*8.5%*	38	*7.5%*	32	*6.3%*	32	*6.3%*
7	32	*6.3%*	22	*4.3%*	15	*3.0%*	15	*3.0%*	15	*3.0%*
8	15	*3.0%*	10	*2.0%*	10	*2.0%*	9	*1.8%*	9	*1.8%*
9	30	*5.9%*	25	*4.9%*	18	*3.6%*	17	*3.4%*	17	*3.4%*

Table 3.6: Absolute number of groups (unique gestures) by gesture length for $n \in \{0,1,2\}$ and their percentage of the dataset. The data indicates that gesture with four cells are exceptionally similar while gesture with six or eight cells show most diversity.

3.5.4 Discussion and Implications

The analysis confirmed that users indeed follow predictable strategies when selecting grid-based unlock gestures. We found that user-defined patterns are usually short (avg. 5 cells), start on the top left of the matrix and end on the bottom right. In addition to these basic results, the application of the proposed similarity metric revealed interesting insights which will be discussed in this Section.

Length is an Important Security Feature

The analysis of user-defined unlock gestures revealed that primarily short gestures are very similar. For example, we found that 21% of all patterns with the length of four could be traced back to simple L⁝-shapes. Overall, the practical pattern space for such short gestures was reduced to 17% of its actual size. This is a serious security issue as over 50% of the participants in this group chose gestures of this length. When users decided to use more cells, the resulting gestures were less similar. We therefore conclude that (similar to alphanumeric passwords) length is a fundamental security measure. However, the results revealed that gesture length cannot be used as a *linear* security feature as the maximum diversity was found for six and eight cells but dropped for patterns using seven cells or the whole grid (length = 9). The manual inspection of such gestures indicated that this phenomenon results from the users' preferences for specific shapes. For example, gestures with seven cells were mainly based on the Z⁚-shape. The use of six or eight cells does not lead to such prominent geometrical shapes.

But There is More to Strength than Length

The discussion of gesture length indicated that the human interest in geometric properties is a strong influencing factor. The proposed metric is the first to reflect such special interest as it considers similarity to other patterns as more important than specific properties of their composition. We argue that specific composition aspects do not necessary increase guessing security as most users prefer the same forms. For example, a minimum length requirement does not increase security, if most users opt for the ⊔⊓-shape. Our metric revealed that knowing a single popular gesture, namely the center of the largest group in the dataset, would have been enough to deduce more than a tenth of the whole dataset by applying only two simple transformations.

This knowledge renders patterns much more susceptible to informed guessing attacks than what is usually anticipated. In addition, the metric is able to identify popular shapes which are not necessarily found in the top ranks of unique patterns but still cluster a significant portion of the pattern space. We thus argue that pattern strength is better assessed by measures that consider human factors, such as geometric similarity, compared to measures that are purely based on obvious properties of pattern composition, such as their length. We assume that one potential solution might be to implicitly manipulate the users' geometric preferences during the selection process (e.g., through user interface design). If such manipulation would be randomly applied, it could increase the overall diversity.

Popular Gestures are Easy to Guess and Easy to Observe

Overall, the analysis confirmed that most users prefer short patterns based on simple shapes. Thus, the selection of Android unlock patterns is even more restricted than previous work assumed. While the preferred use of L⁚-shapes and Z⁚-shapes was already reported, we found that most patterns which seem unique at first glance are close relatives of these shapes. Therefore, we conclude that most gestures are easy to guess considering the top gestures and up to two simple transformations.

In addition, the application of our prediction model (see Section 3.4) indicates that most user-defined gestures are also easy to observe. 52% of the users used four cells without special moves. Assuming that strokes are visualized, the model predicts an observability quotient of *91*. This indicates with almost absolute certainty that an observation attack is successful. Indeed, the data reveals that 92.5% of the four-cell gestures were successfully shoulder surfed during the user study. We conclude that most real-life gestures are both easy to observe and easy to guess. If attackers can combine knowledge from both attacks, they are very likely to succeed.

3.5.5 Limitations

The metric reflects that users are guided by geometric properties when creating patterns. However, we currently apply a simple gesture distance, namely counting geometric transformations. Furthermore, we do not know how different geometric transformations compare to one another in terms of the users' *perceived similarity*. That is, transformations should possibly be counted with different relative weights not all contributing to the total distance with the same weight, as assumed here. Such weights need to be determined by future studies. In addition, pattern groups found with the greedy approach may not match the global optimum. However, as the main insights are derived from analyses of the largest groups and the most popular patterns, we expect them to be rather stable.

Considering the evaluation strategy, not all confounding factors could be ruled out. The data set was collected via Amazon Mechanical Turk and thus mostly collected from young US citizens. Therefore, the data set might not be representative of other age groups and cultures. In addition, since we could not collect real-life gestures, we based the analysis on gestures which were collected during a simulated real-life situation. We like to note that real-life gestures might slightly vary dependent from the used context. However, since the basic statistics matched the findings of prior work, we are confident that most of the participants contributed patterns which they would also have used in a real situation.

3.5.6 Summary

In this Section, we investigated the similarity of grid-based unlock gestures. For this purpose, we proposed a novel metric and applied it to a corpus of 506 user-defined unlock gestures. The analysis revealed that most users base their secrets on a small set of very similar shapes. Considering all gestures with a distance $n \leq 2$ as duplicates reduced the practical gesture space of unlock patterns in the data set by approximately 66%. In addition, we found that most gestures were short (avg. 5 cells) and comprised only straight lines. The results indicate that user-selected unlock gestures are very predictable and prone to dictionary attacks. We therefore conclude that solutions to make user-selections more diverse are required.

The insights presented in this Section are important for motivated users who want to strengthen their choice of unlock gestures. Furthermore, this work provided both the motivation and the tools to research potential countermeasures against predictable pattern selection. Chapter 4 will discuss such strategies and propose systems which have the potential to diversify gesture choice. The systems will be evaluated based on the similarity metric presented in this Section.

3.6 Result Aggregation and Implication

We presented four research projects which explored the problem space of gesture-based authentication on mobile device. The performed user studies provided valuable insight into real-world aspects and contributed to an in-depth understanding of current systems. In this Section, we summarize the main results and draw implications for future designs. Please refer to the respective sections for a more detailed discussion.

3.6.1 Lessons Learned

In Chapter 2, we identified important open questions regarding the real-world unlock behavior and the real-world performance of current unlock concepts. In this Chapter, we systematically filled most the identified gaps by defining and exploring the problem space of gesture-based authentication.

We learned that device unlocks are performed at high frequency but usage sessions are usually short. In addition, field observations revealed that risks are seldom received and users are not willing to invest additional effort for additional protection. We concluded that this specific unlock behavior in combination with low risk perception significantly increases the importance of *efficiency*. In this context, we also found that *perceived efficiency* is actually more important than quantitatively measured values. However, results indicated that, irrespective of the used authentication concept, unlock times up to four seconds were still acceptable for most users. At the same time, the results indicated that high unlock frequencies reduce *memorability* problems since the user's memory is constantly refreshed.

Furthermore, the investigation confirmed the importance of *effectiveness*. However, we found that effectiveness cannot be equated with success rate. This is especially true whenever deliberately failing is more efficient than error correction. In this connection, we found that current unlock gestures outperform numeric approaches. Even if the quantitative assessment indicated higher error rates, we found that the specific error handling of the gesture approach was rated significantly better. This indicates that the mobile context renders fast error recovery more important than error correction.

Even though gestures did not always outperform PIN, we found that they were *perceived* to work equally well. In addition, participants stated that using gestures feels good and most users were generally in favor of the gesture-based concepts. In terms of usability and *likeability*, this indicates that current gesture-based concepts are already close to the goal state. At the same time, we found that using current unlock concepts opens serious security holes. The analysis indicated that users *select* gestures from a very *limited* pool of shapes. In addition, we found that most real-world gestures are easy to *observe*. Even though we found that such risks are seldom perceived and security can be partly improved through gesture selection, the current state is far away from the goal state and thus, we argue that such problems need to be solved by future designs.

Finally, the Chapter contributed two strength models which help to assess the *practical security* of a given gesture. A prediction model helps to quantify *observation* risk, a similarity metric considers the human preferences for geometric properties and helps to assess *guessability*. The contribution of these models is twofold: Firstly, the metrics can help researchers to further investigate human selection behavior and to benchmark potential countermeasures. Secondly, the outcome can immediately help motivated users to select more secure gestures.

3.6.2 Implications for Future Designs

The analysis of the current state of gesture-based authentication mechanisms has important implications for future designs. In Chapter 4, we will explore the design space and investigate how gesture-based authentication methods can be designed in a way that improves the current state and brings it closer to the goal state. For this purpose, we have to prioritize the different aspects of the problem space according to both the distance between current state and goal state and the importance in the users' everyday life.

Usability

While the ultimate goal in terms of usability must be full accessibility, 100% effectiveness and zero effort, we conclude that the usability of the next generation of gesture-based authentication concepts must be at least comparable to current solutions. Current solutions are already widely accepted and the current state is close to the goal state. Therefore, when developing novel concepts, good *efficiency* must be the main goal. We argue that, independent from provided security benefits, bad performance is a criterion for exclusion for any mobile unlock mechanism. The same is true for *effectiveness*. In this connection, quick error recovery shall be prioritized. Concerning *memorability*, we conclude that the factor is less important in the context of mobile unlock systems. However, common limits of human memory need to be considered. That is, adequate *theoretical security* shall be reached within a feasible number of memorable chunks (i.e., gestures) [17]. Furthermore, since the results render *perception* aspects extremely important, we conclude that novel concepts shall be designed in a way that increases likeability. This also implies that user perception and likeability needs to be individually assessed during evaluations. Finally, all concepts shall be as *accessible* as current solutions. That is, all solutions shall be tailored to the mobile context (e.g., support one-handed interaction) and supported by current devices or shortly available solutions.

Security

The field evaluation revealed that security risks are hardly perceived by users. At the same time, the analysis of current gestures showed that current concepts are far away from the goal state. We argue that, even if threats are rarely perceived, the next generation of gesture-based authentication concepts must provide improved security. However, the results of the field study indicate that users are not willing to accept reduced usability for increased security. Concerning *theoretical security* aspects, we conclude that current concepts already provide adequate protection. Therefore, novel concepts need to provide at least the same level of security. That is, a minimum of 10,000 secrets should be provided and encrypted storage of such secrets must be possible. In contrast, the *practical security* of current unlock mechanisms shows major deficits. Researchers need to consider human preferences (e.g., for geometric properties) for specific gestures and develop concepts which have the potential to implicitly break such habits and make gesture selection less *guessable*. Furthermore, practical security must be improved in terms of *observation resistance*. Since the field evaluation indicated that users are often casually observed but advanced shoulder surfing attacks are seldom, we conclude that the protection from such casual attacks should be prioritized. Nevertheless, the ultimate goal must be perfect protection, even from camera attacks. The third serious threat is *smudge*. Even if smudge attacks were not investigated in this Chapter, Chapter 2 indicated that current concepts are vulnerable. Therefore, the design of *smudge resistant* unlock gestures must be a goal for future concepts. Finally, Chapter 2 indicated that *divulgation resistance* is less a problem in connection with touch gestures. Still, this factor should not be neglected and novel concepts should achieve at least similar protection as current solutions.

Chapter 4

Exploring The Design Space of Gesture-based Authentication

It's not just what it looks like and feels like.
Design is how it works.

– Steve Jobs, Chairman and CEO of Apple Inc. (2003) –

Based on the insights of Chapter 3, this Chapter addresses the main problems of established gesture-based authentication mechanisms: Smudge Attacks, Observation Attacks and Guessability. For this purpose, novel concepts have been developed and evaluated in the lab and in the field. The results of such studies illustrate the impact of different design factors and contribute to the understanding of gesture-based authentication on mobile devices.

Section 4 illustrates the different aspects of the *design space* and gives an overview of the research covered in this Chapter. Section 4.2 addresses the threat of smudge attacks and investigates the utility of randomization to develop more resistant authentication mechanisms. The evaluation shows that randomized interfaces can be both efficient and effective. In Section 4.3, we focus on observation attacks and present two projects which especially investigate the impact of visual cues. The evaluation shows that clever interaction design allows building very efficient authentication methods which provide observation-resistance on demand. Section 4.4 systematically investigates output elements and presents two concepts which utilize static guidance to implicitly influence gesture selection. The results show that minor changes in the interface of established authentication mechanisms can have a high impact on practical security. Finally, Section 4.5 summarizes the main findings and discusses the impact of different design decisions.

4.1 The Design Space

Chapter 2 illustrated various ways to design graphical gesture-based authentication mechanisms. Some concepts utilize gestures to activate visual elements represented on the screen [18], other systems use the gesture itself as secret information [227]. In this Section, we define the design space of graphical gesture-based authentication on mobile devices. The first part presents the main design factors and points out potential influences on the usability and the security of an authentication system. The second part gives an overview of the research projects which were performed to explore the defined design space.

4.1.1 Definition

We like to note that the presented list of design factors is not exhaustive. We rather take a simplified view which allows a systematic evaluation of the most important factors. Although parts of the design space may apply to other device classes and other interaction concepts as well, the presented design space is specifically defined to match the combination of touch gestures and touch-based mobile devices. This especially means that the definition does not consider multimodal concepts (e.g., sound or haptic cues). Inspired by Schaub et al. [214], we define four main categories: *Input Elements*, *Output Elements*, *Element Arrangement* and *Interaction Style*. Figure 4.1[1] gives an overview of the design space.

Input Elements

Input elements represent all active areas which can be used to enter data. Therefore, input elements are essential for any touch-based authentication system. Input elements can differ in size, shape, texture, text and number. In general, the design should comply with standard design guidelines for mobile devices. First of all, it should be optimized for finger-based interaction and input elements should be reachable with one hand. For example, Google recommends a minimum size of $48 * 48$ dp (density-independent pixels) for touch targets and at least 8dp between targets[2]. Similar values are recommended by Apple [3]. In addition to the compliance with general design guidelines, we identified the *representation* and the *reusability* of the input elements as important design factors.

Representation The visual representation of the input element is distinguished into *none*, *abstract* and *concrete*. *None* means that there are no visual representations of the input elements. This might be the case whenever the input element fits the whole screen or whenever the input is guided by additional output elements. If input elements are visually represented,

[1] All figures in Section 4.1.1 are based on flat icon designed by Freepik http://www.freepik.com, licensed under CC BY 3.0

[2] https://www.google.com/design/spec/layout/metrics-keylines.html – accessed: 2016/02/17

[3] https://developer.apple.com/library/ios/documentation/UserExperience/Conceptual/MobileHIG/LayoutandAppearance.html – accessed: 2016/02/17.

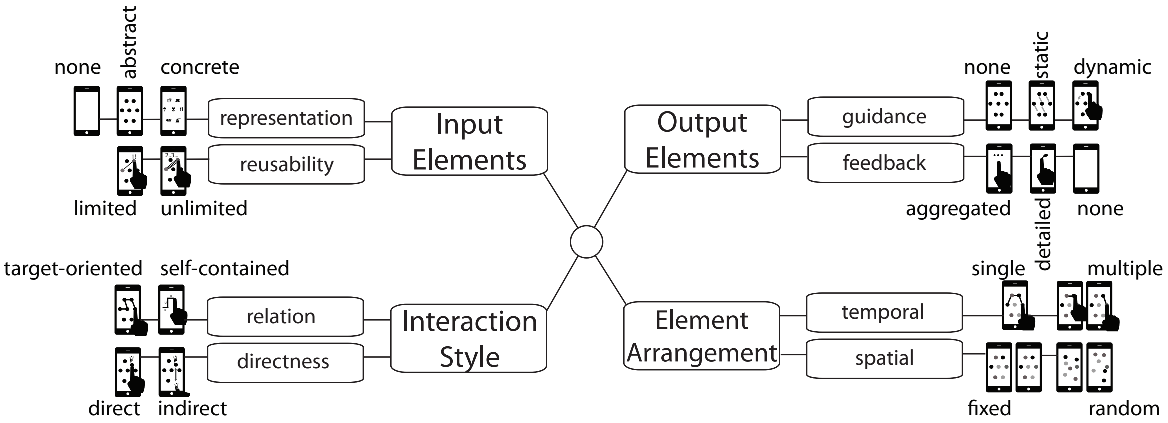


Figure 4.1: The Figure illustrates the most important aspects of the design space of graphical gesture-based authentication mechanisms on mobile devices.

they can be abstract or concrete. *Abstract* representations imply all shapes and colors without physical referents. In contrast, *concrete* input elements represent physical objects or well-known symbols like letters and digits. Figure 4.2 illustrates the different representations. While the abstract input elements are represented by black circles, the concrete input elements represent food and kitchen objects.

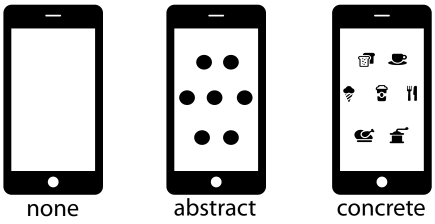


Figure 4.2: Visual representations of input elements can be abstract or concrete.

We hypothesize that the representation of input objects influences both usability and security. While omitting visual representations can increase observation resistance, the use of visual representations is likely to increase usability. Section 4.2 will show that concrete representations can improve the user experience and support story-based memorization. Section 4.3 will present an authentication concept which supports eyes-free interaction and thus completely forgoes visual representations.

Example: The *Android pattern unlock* provides nine *abstract* input elements which are usually represented by dots.

Reusability The reusability of input elements defines how often an input element can be selected (activated) during enrollment or authentication. We distinguish between limited and unlimited reusability. *Limited* reusability represents cases in which input elements become unavailable after a certain number of activations. In contrast, unlimited reusability describes the case where the number of activations is user-defined. Reusability is likely to affect the composition of the secret as well as efficiency and effectiveness. Section 3.3 already indicated that reusable elements increase the risk of unwanted activation. In contrast, limiting the reusability of input elements usually reduces the theoretical password space. Figure 4.3 illustrates both cases. The *limited* interfaces allows to activate input elements once while the right interface allows unlimited activation.

Example: The reusability of the *Android pattern unlock* is limited to one activation per cell.

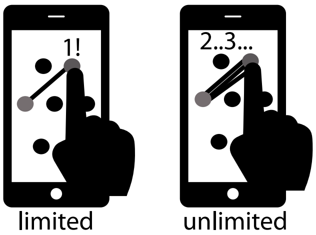


Figure 4.3: Input elements can either support unlimited reuse or limit the number of activations.

Output Elements

Output elements represent all visual elements which are not interactive. Schaub et al. [214] defined such elements as visual cues. Output elements can display any kind of information (e.g., text). In contrast to input elements, the existence of output elements is optional. Similar to input elements, the number and the visual appearance of output elements play an important role for usability and security. In general, the presented information needs to be easy to understand and optimized for fast processing [253]. We categorize output elements according to their purpose: *Feedback* and *guidance*.

Guidance Guiding elements are used to guide the user's actions or to influence her decisions. This especially implies that the output is bound to subsequent interactions. If guidance is available, we distinguish between static and dynamic elements. *Static* guidance describes output elements which are independent from user interaction while *dynamic* guidance is bound to user input or other events (e.g., unlock events). Figure 4.4 illustrates the characteristics of guiding output elements. While the static elements illustrate general options for the next target, the dynamic elements may recommend one specific target depending on previous interactions.

Guidance can generally support usability as it simplifies the interaction. In addition, we will learn how dynamic guidance can be used to improve the practical security of an authentication system. Section 4.2 illustrates the importance of guidance in connection with randomized interfaces. In Section 4.3, we will demonstrate how dynamic guidance can be used to protect input from observation attacks. Section 4.4 will present several concepts which exploit static guidance to diversify the selection of gestures.

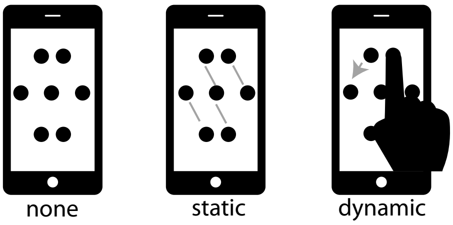


Figure 4.4: If guidance is provided, it can be distinguished into static and dynamic elements.

Example: The *Android pattern unlock* does not provide guidance. However, it could be implemented by highlighting the cells which are reachable from a certain position.

Feedback Feedback is important for any user interface to indicate the state of elements or to confirm that touch input was received. In contrast to guiding elements, feedback elements are bound to previous user input or an earlier event. We distinguish feedback according to its information content: feedback is either aggregated or detailed. In addition, some systems may also forgo visual feedback. *Aggregated* feedback informs the user in an abstract way, while *detailed* feedback gives concrete information.

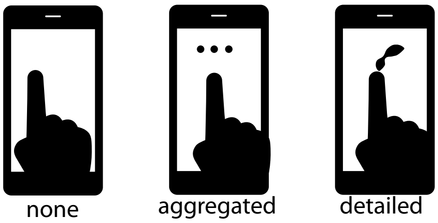


Figure 4.5: If feedback is provided, it can be given in an aggregated or in a detailed manner.

Figure 4.5 illustrates the different options. The example shows that aggregated feedback may acknowledge user input by displaying binary symbols (received versus not received). In contrast, the example for detailed feedback visualizes the entered gesture. If an error occurs, the user would be able to identify the exact source of error. Therefore, detailed input is likely to increase usability and may also increase memorability in a way that the performed gesture is visually perceived. On the other hand, detailed feedback may give away information to potential observers. Section 4.2 presents several prototypes which make use of detailed and aggregated feedback. In Section 4.3, we will deliberately reduce the detail level of feedback to increase observation resistance.

Example: The *Android pattern unlock* gives detailed feedback per default. Gestures are visualized by highlighting both the drawn path and the activated cells. Errors are represented by visualizing the entered gesture in a different color (usually red). While the main feedback can be switched off, error highlighting is always activated.

Element Arrangement

Element arrangement is an important factor which directly influences the usability and can influence the security. Input and output elements can be arranged in different ways. Following Schaub et al. [214], we differentiate between *spatial* and *temporal* arrangements.

Spatial Spatial arrangement describes how elements are positioned on screen. We distinguish fixed and random arrangements. If an authentication system provides a fixed layout, all elements appear at a predictable (usually the same) position. In contrast, a randomized layout positions elements at unpredictable places. While fixed spatial arrangements support motor memory and thus usually increase efficiency and effectiveness, randomized layouts may increase practical security. Figure 4.6 illustrates both arrangements.

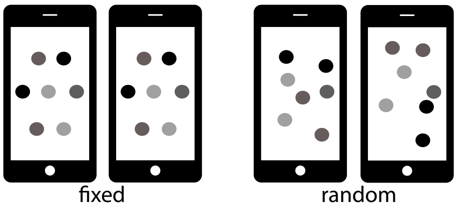


Figure 4.6: Input elements and output elements are either presented using a fixed spatial arrangement or are positioned using a randomized layout.

Section 4.2 will present several concepts which illustrate how randomized layouts can be used to prevent smudge attacks. In addition, Section 4.3 and Section 4.4 will present concepts which rely on randomized output elements. We will learn how randomized output can increase both observation resistance and the practical password space.

Example: The *Android pattern unlock* is based on a fixed spatial arrangement of both input elements and output elements.

Temporal The temporal arrangement depicts both, the number of challenges and the continuity of the user interaction. According to Schaub et al. [214], we distinguish between *single challenges* and *multiple challenges*. Single-challenge-arrangements are often based on the input of one continuous gesture. In contrast, concepts which are based on multiple challenges always require multiple discrete gestures. Figure 4.7 illustrates both configurations. While the concept on the left requires a single challenge, namely the input of a single gesture, the concept on the right requests multiple challenges.

Section 4.2 investigates the effects of single-challenge arrangements and multi-challenge arrangements in combination with randomized spatial layouts. We will learn that the temporal arrangement can have a significant impact on the perceived efficiency. In Section 4.3, we discuss how multiple challenges can be used to protect gesture input from observation.

Example: The *Android pattern unlock* is based on a single challenge (i.e., one continuous gesture).

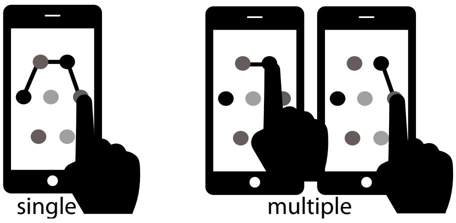


Figure 4.7: Authentication concepts can be based on a single challenge or on multiple challenges which need to be accomplish in sequential order.

Interaction Style

While gesture-based interaction is a precondition for all concepts in the scope of this work, the interaction style varies. We categorize the interaction style according to the *relation* of the gestures and other elements (targets) and according to the *directness* of the interaction.

Relation Relation describes if the information content of a gesture is self-explanatory or if the information is derived from other elements on the screen. It therefore describes the relation of the gesture and other elements. If the performance of a gesture is not bound to any targets, we call it *self-contained*. Gestures which are self-contained are often based on relative movements. In contrast, gestures which are not self-contained are called *target-oriented* as the information is depending on the activated target. Figure 4.8 gives an example for each gesture type. On the left, the gesture is target-oriented as it is mainly guided by the elements on screen. The concept on the right illustrates a self-contained gesture which is defined by relative shifts in direction (i.e., right,up,right).

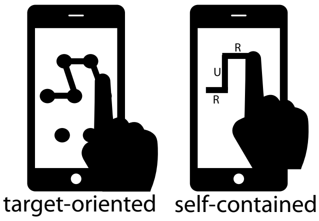


Figure 4.8: Target-oriented gestures are described by the activation of specific input elements while self-contained gestures are often described by relative movements.

Section 4.2 will present several target-oriented authentication concepts. The evaluation will show that the way of target binding plays a significant role for likeability and performance. Section 4.3 will illustrate a concept that uses self-contained gestures to enable eyes-free input and prevent observation attacks.

Example: The *Android pattern unlock* uses target-oriented gestures. The gesture becomes ineffective if the set of activated targets differs from the set of expected targets.

Directness Beside the relation of the gesture and the input elements, the relation of input elements and output elements plays a significant role. *Direct* gesture input represents the normal case. However, gesture input may also be performed independently from the representation of the input elements. We hypothesize that such *indirect* gesture input can increase practical security in terms of smudge attacks or observation attacks. Figure 4.9 illustrates a target-oriented authentication concept. On the left, the gesture is performed directly on the respective representation of the input element. On the right, the gesture is performed indirectly as the input takes place at a position which differs from the position of the visual element.

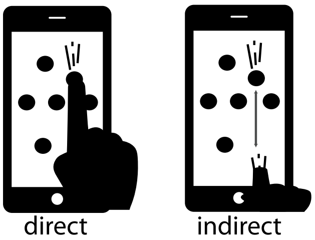


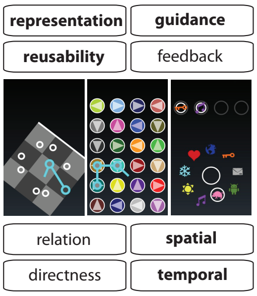
Figure 4.9: Gestures can be performed directly on the target element or indirectly. Indirect gestures are described by input areas which differ from the visual representation of the respective target elements.

Section 4.2 will present concepts that are based on direct gesture input, while Section 4.3 presents two concepts that exploit indirect gesture input to increase observation resistance.

Example: The *Android pattern unlock* is based on direct gesture input as the touch-coordinates need to match the visual representations of the input elements for activation.

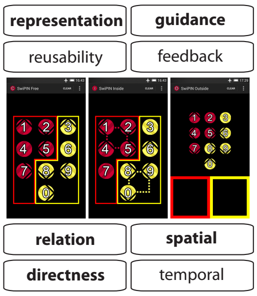
4.1.2 Overview

This Chapter presents three research projects which were conducted to explore the design space of graphical gesture-based authentication on mobile devices. The overall goal of all projects was to find the right design to bring the current state of gesture-based authentication closer to the previously identified goal state [8]. The projects are presented in three sections. In each Section, we present and discuss the development process, the concepts and the results of laboratory experiments and field studies. Each concept was implemented as interactive prototype and thoroughly studied. The following gives an overview of the research covered in this Chapter.



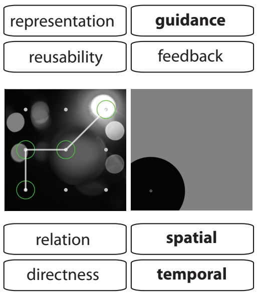
On Preventing Smudge Attacks

Section 4.2 explores the design space in the light of smudge attacks. We present several authentication mechanisms which were designed in a way that smudge traces are hard to interpret. For this purpose, we specifically focus on the effects of randomized *spatial* and *temporal* arrangements. In addition, different aspects of *reusability*, *guidance* and *feedback* will be discussed.



On Preventing Observation Attacks

Section 4.3 explores the design space with the aim of developing observation-resistant authentication mechanisms which are fast and easy to use. We present two concepts which make use of *direct* and *indirect* gesture input. Both systems allow the user to adjust the security level according to the current situation. We make use of *target-oriented* gestures and *self-contained* gestures and will learn how *dynamic guidance* can be used to prevent shoulder surfing.



On Increasing the Practical Password Space

Section 4.4 differs in various ways from the other two sections as the presented projects aim to diversify gesture selection. The two presented projects focus on the enrollment phase and not on the authentication procedure itself. We will investigate the effects of implicit *guidance* and learn that the *spatial arrangement* of visual cues can significantly influence gesture selection. The resulting gesture sets will be analyzed based on the metric presented in Section 3.5.

4.2 On Preventing Smudge Attacks

This Section explores the design space in terms of *smudge attacks*. As revealed by Aviv et al. [18] in 2010, current graphical gesture-based concepts are very much prone to such attacks. We present eight instances derived from three different concept classes. All concepts are particularly secure against smudge attacks as interaction leaves smudge traces which are hard to interpret. The results of this Section are based on an iterative design process that involved low-fidelity and high-fidelity prototyping and evaluations in the lab and in the field.

This Section sheds light on the following main *research questions*:

RQ1 How can *randomized spatial arrangements* be utilized to build smudge-attack resilient but yet usable authentication mechanism for mobile devices?

RQ2 How does *randomization* influence user acceptance and user perception and which are the important *design factors*?

RQ3 How does *randomization* influence effectiveness and efficiency and which are the important *design factors*?

RQ4 How does *randomization* influence learnability and memorability and which are the important *design factors*?

The evaluation is based on three *lab studies* (n = 54) and on a *field study* (n = 18). While this Section focuses on the rather limited threat model of smudge attacks, it allows a much broader view on the interplay of the *design space* and the *problem space*.

While all concepts are based on *direct target-oriented* gestures, we will learn how *randomized spatial arrangements* of input elements can protect touch gestures from smudge attacks (*RQ1*). At the same time, the evaluation will show that the *temporal arrangement* of such randomized interfaces plays a vital role for user acceptance (*RQ2*). While *multiple* challenges tend to increase the input time, they are perceived more efficient. In contrast, providing *single* randomized challenges led to reduced user acceptance and lower perceived efficiency even though these concepts were measurably faster (*RQ3*). In addition, the Section will provide valuable insights into the design of input elements concerning *reusability* and *representation* and show that *guiding* output elements play a significant role for both usability and security. Finally, the field evaluation will show that randomized concepts indeed allow learning effects (*RQ4*).

Parts of this Section are based on von Zezschwitz, E., Koslow, A., De Luca, A., & Hussmann, H. (2013, March). Making graphic-based authentication secure against smudge attacks. In Proceedings of the IUI'2013 [268]. In addition, parts are based on a bachelor thesis by Alexander Kehr [144] which was carried out under my constant supervision. Please refer to the beginning of this thesis for a detailed statement of collaboration.

4.2.1 Research Context and Motivation

Since related work is already well covered in Section 2.3.2, this Section focuses on a short overview. As mentioned, Aviv et al. [18] were the first to show that Android unlock gestures are prone to smudge attacks. The authors presented the results of a lab study where 68% of the entered gestures could be successfully deduced from oily residues left on the touchscreen.

In the following years, the work motivated several researchers to develop concepts which are specifically designed to reduce the risk of smudge attacks. Two of the first concepts were Vertical PIN and WhisperCore [4] which both introduced additional wiping tasks to blur the actual smudge traces. Oakley and Bianchi [195] presented a modification of Android unlock patterns which utilized multi touch. The authors assumed that the overlapping strokes could reduce the risk of smudge attacks.

In addition to the work presented here, a few other concepts were recently published which utilize randomized spatial arrangements to prevent smudge attacks. SmudgeSafe by Schneegass et al. [218] uses a randomly applied set of geometric transformations (including rotation) to alter smudge traces. The transformations are applied to background images which serve as representations of input elements. SwiPass by Kosugi et al. [158] randomly displays images taken by the user. As swipe directions are depending on the relative age of the image, they dynamically change and smudge traces are hard to interpret. Finally, Amruth and Praveen [7] proposed several concepts which are similar to WhisperCore [4] and multi-touch gestures [195].

In addition to these randomized concepts, other approaches might also achieve smudge attack resistance. TinyLock [161] minimizes the interaction area by minimizing the $3x3$ grid of Android unlock gestures. As a consequence, smudge traces overlap and become hard to interpret. Chiang and Chiasson [52] propose multi-layered gesture input. Using multiple layers may result in overlapping smudge traces. In addition, it may be hard to deduce the actual layer a gesture was performed in. Glassunlock [281] combines a randomized PIN layout with guidance through smartglasses. The combination makes authentication secure against both smudge attacks and observation attacks. Finally, smudge attack resilience can be achieved by adding implicit authentication layers using behavioral biometrics [69].

4.2.2 Threat Model

Mobile interaction goes far beyond authentication. As a consequence, smudge traces are typically constantly modified and visual cues become harder to interpret with every additional input. However, our threat model assumes a best case scenario for an attacker who tries to perform a smudge attack. According to the threat model, the attacker has temporary access to the mobile device. For example, the mobile device was left unattended at am office desk. The attacker uses this opportunity to clean the surface of the touch screen. As the user returns she does not recognize the malicious activity and authenticates to check an incoming text message. Besides reading the message, no interaction takes place. Next time the

		Chessboard	**Compass**	**Connect Four**	**Dial**	**Marbles**	**Pattern Rotation**
Input Elements	*Representation*	abstract	abstract	abstract	abstract	abstract or concrete*	abstract
	Reusability	limited	unlimited	limited	unlimited	unlimited or limited*	limited
Output Elements	*Guidance*	static	static	dynamic	static	static	static
	Feedback	detailed	detailed	detailed	detailed	detailed	detailed
Interaction Style	*Relation*	target-oriented	target-oriented	target-oriented	target-oriented	target-oriented	target-oriented
	Directness	direct	direct	direct	direct	direct	direct
Element Arrangement	*Temporal*	single	single	single	multiple	multiple	single
	Spatial	random	random	random	random	random	random

Table 4.1: All concepts are based on randomized layouts, but the specific kind of randomization is different. *For the sake of brevity, modifications of *Marbles* were grouped.

user leaves the mobile device unattended, the attacker gets again in possession of the device. Since smudge traces are clearly silhouetted against the clean touchscreen and the device is protected by a fixed grid-based unlock scheme (e.g., Android gesture unlock), the attacker can easily deduce the entered secret and gains full access to the device.

4.2.3 Concept Overview

The following concepts were designed to leave smudge traces which are not easy to interpret. However, as preliminary user studies indicated early that not all candidate concepts are feasible for daily use, not all concepts ran through the whole development process. Table 4.1 gives an overview of the different approaches.

Consecutive Blurred Smudge Traces

Two concepts were mainly based on the idea of blurring smudge traces. Therefore, the interaction is designed in a way that gestures usually pass the same screen location multiple times. While the idea seemed promising at first, both concepts were rejected after low-fidelity prototyping.

Compass is a drawmetric concept (see Figure 4.10, left). In contrast to Android unlock gestures, input elements are arranged in circular order. We assume that the circular layout will lead to an increased interference of smudge traces and hamper interpretation. In addition, smudge attack resistance is increased by randomly rotating the circle of input elements (viewport). However, analogous to a real compass, the internal order of the elements stays the same. The current orientation of the compass is indicated by static output elements: The arrow and the initials of three cardinal points. A gesture results from connecting several input elements, while the same input element can be visited multiple times.

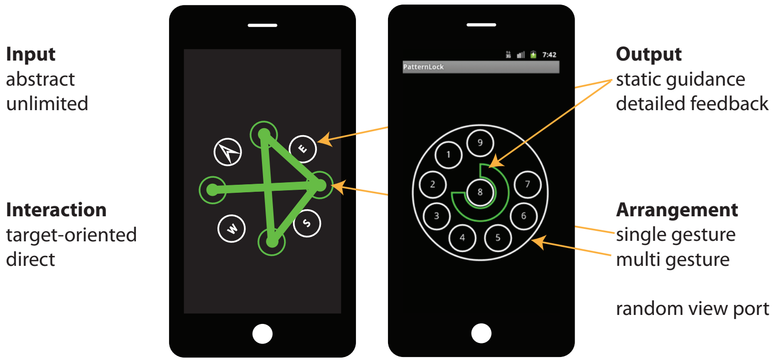


Figure 4.10: Both concepts are based on circular arrangements. With *Compass* (left) users connect several input elements to authenticate. *Dial* (right) works analogous to a dial plate as input elements are successively dragged to the center of the screen.

Dial is a numeric concept which uses gestures for input (see Figure 4.10, right). Similar to *Compass*, input elements are arranged in circular order. Analogous to an old-fashioned dial plate, the internal order of the input elements stays the same. In addition, the viewport is rotated within a range of -45 degree and +45 degree. A password consists of an arbitrary sequence of digits. An authentication consists of multiple gestures. Each gesture is described by dragging one of the input elements (digits) into the center of the screen. The dragging path is specified by static output elements: Elements have to be dragged within the margins and through the opening under the element “1”. As a consequence, smudge traces are consecutively blurred.

Randomly Rotated Viewports

Two concepts are based on viewport rotations. Such concepts improve smudge resistance by increasing the degree of freedom of a entered gesture. Nevertheless, if smudge traces are clearly visible, attackers still have a good chance to guess the right gesture. While both concepts were rejected during the development process, *Pattern Rotation* was the basis for *Connect Four*, a promising concept which will be described in the next Section.

Pattern Rotation is based on the Android pattern unlock (see Figure 4.11, left). However, the matrix is randomly rotated on the screen and translated along the y-Axis. The current orientation is indicated by an arrow. We tested two different versions: A 90-degree version with four different orientations (degrees of freedom) and a 360-degree version which allowed arbitrary orientations. If the authentication takes place on a clean screen, the entered gesture can still be derived. For example, the attacker would have a chance of 1:4 to guess the right orientation whenever the 90-degree version was used. However, we assume that the varying rotation of the matrix would usually blur smudge traces.

Chessboard is a modification of the pattern rotation scheme which was developed in the second design cycle (see Figure 4.11, right). In addition to the viewport rotation known from

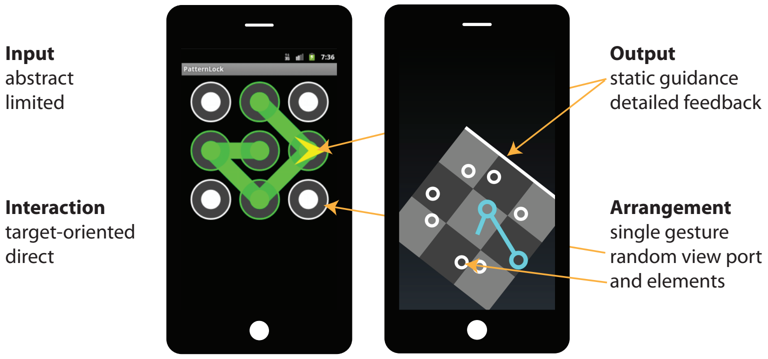


Figure 4.11: *Pattern Rotation* (left) is based on the $3*3$ matrix layout which is randomly rotated and translated across the y-Axis. *Chessboard* (right) is a modification of *Pattern Rotation* which additionally implements a random spatial arrangement of input elements.

the pattern rotation system, the concept utilizes randomly floating input elements. That is, active input elements are represented by circles and float within the borders of the squares (cells). The orientation of the matrix is indicated by a white line. We assume that implementing a fuzzy arrangement of the nine input elements leads to smudge traces which are harder to interpret. However, even if the evaluation showed that this assumption is right, the *Chessboard* layout had to be rejected due to usability drawbacks.

Randomized Input Elements

The two most promising concepts are based on a fully randomized spatial arrangement of input elements. *Marbles* was developed in the first development cycle and constantly improved; *Connect Four* evolved as a modification of *Pattern Rotation*. Both concepts ran through all development stages and were evaluated in the lab and in the field.

Marbles is based on the randomly distributed input elements (see Figure 4.12). Over the course of the project, three different versions were designed. The original *Marbles* concept is illustrated in Figure 4.12, left. A secret consists of an arbitrary sequence of differently colored input elements while each element can be selected multiple times. To authenticate, input elements are dragged into the center of the screen. After a marble is logged, it immediately reappears on its prior position. The position of the input elements is fixed during one authentication but changes for each new attempt. Therefore, smudge traces reveal no information on the used secret. In the second development cycle, we designed *Marbles Story* (see Figure 4.12, center) which provides the same functionality as the original scheme but uses concrete representations of input elements.

Marbles Gap is a modification of the original scheme which provides a different spatial arrangement. In contrast to a circular arrangement, the screen is divided into three sections. The top and the bottom Section provide the input elements while the Section in the center is called “gap”. To authenticate, marbles need to be dragged into the gap. As indicated in

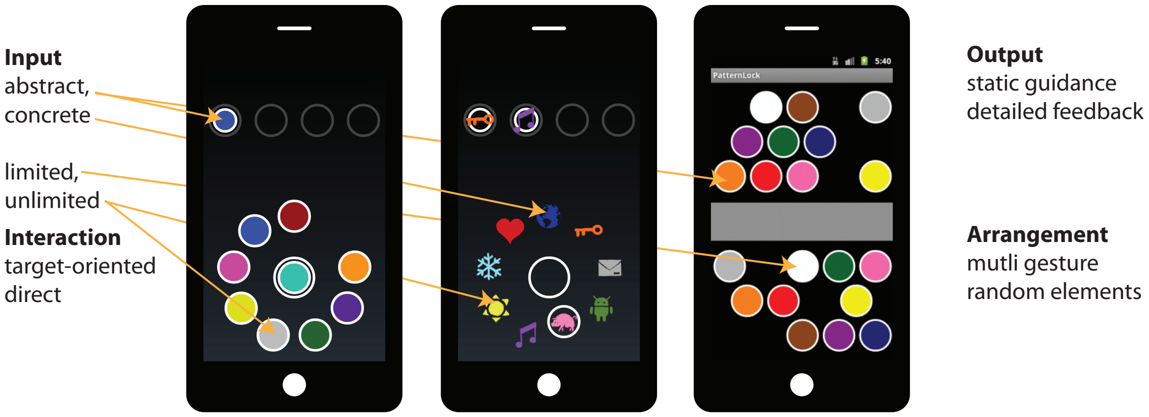


Figure 4.12: The original *Marbles* concept (left) and *Marbles Story* (center) are based on the arbitrary selection of circularly arranged input elements. In contrast, *Marbles Gap* (right) uses the metaphor of a gap to limit reusability of input elements.

Figure 4.12 (right), each color is represented by two input elements (one in each segment). Marbles of the same color are equivalent. In contrast to the original scheme, the reusability of input elements is limited as marbles disappear as soon as they were dragged into the gap. As a consequence, secrets can comprise a limited number of identically colored marbles (i.e., two). Finally, choosing the same color twice requires more interaction effort (i.e., drag elements from the bottom and the top) than choosing differently colored elements (i.e., drag marbles from one side only). We assume that this interplay of input effort and password diversity is likely to increase the practical password space.

Connect Four is a modification of the *Pattern Rotation* scheme. In contrast to *Pattern Rotation* and *Chessboard*, *Connect Four* is based on a fully randomized arrangement of input elements. As a consequence, *Connect Four* does not rotate the matrix itself but randomizes the orientation of input elements. Input elements are represented by differently colored items. In addition, each input element is tagged with an arrow. The arrows are static guiding elements that randomly indicate one of four directions. Figure 4.13 illustrates the concept during enrollment (left) and during authentication (right). During enrollment, the center of the grid shows a specific color (i.e., green) and a direction (i.e., down). The user selects a gesture following the common Android policy. As a consequence, the secret is a combination of the color information, the orientation and the gesture. To authenticate, the user has to find the center element and perform the gesture in relation to the indicated direction. As a $3*3$ grid is required to perform the gestures, the outer line of elements cannot serve as center (guiding element). Therefore, smudge traces can result from eight different center elements which can indicate four different directions. We assume that this factor shrinks the chance for a successful smudge attack, even if the smudge trace can be interpreted. In addition, the concept provides increased protection from observation attacks as one-time observers do not know which color currently serves as guiding output element.

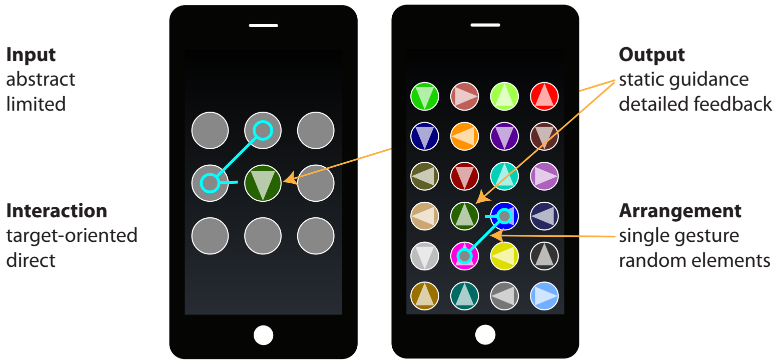


Figure 4.13: *Connect four* is a modification of pattern rotation which is based on randomized input elements and a secret cue. During enrollment, the system displays a specific color and a direction (left). During authentication, the user recognizes the assigned color (right).

4.2.4 Building the Foundation for Smudge-resilient Gesture Input

As indicated by Figure 4.14, the concept development was divided in two phases. The first development cycle built the foundation for smudge-resilient gesture input. It started with an ideation phase which identified the spatial arrangement of input elements as the main design factor. At the end of the ideation phase, we came up with four concept candidates. Next, we performed two user studies which were based on low-fidelity paper prototypes and high-fidelity software prototypes. The results provided valuable insights into the effects of randomized spatial arrangements and built the basis for the second development cycle. The first part of this Section presents the evaluation strategy and the designs of the performed user studies. The second part focuses on the results.

Prototypes and Evaluation Strategy

The first development cycle included a paper prototyping study and a lab study.

Evaluation 1: Paper Prototyping

To get a first impression of user acceptance and to identify basic usability problems, we built paper prototypes and evaluated them in a preliminary user study.

***Design**:* The study was based on a repeated measure within participants design. The independent variables were *system* with five levels (*Marbles*, *Dial*, *Compass*, *Pattern Rotation 90*, *Pattern Rotation 360*) and *password type* with two levels (*given*, *self-selected*). *System* was counterbalanced based on a Latin square design, *password type* was alternated.

***Procedure and Setup**:* Each system was tested with a given secret and with a self-selected secret. The examiner explained the concept and handed the prototype over to the participant. Whenever the participant felt ready, the prototype was used to create a password and to authenticate once. Next, participants rated usability, likeability and perceived security using

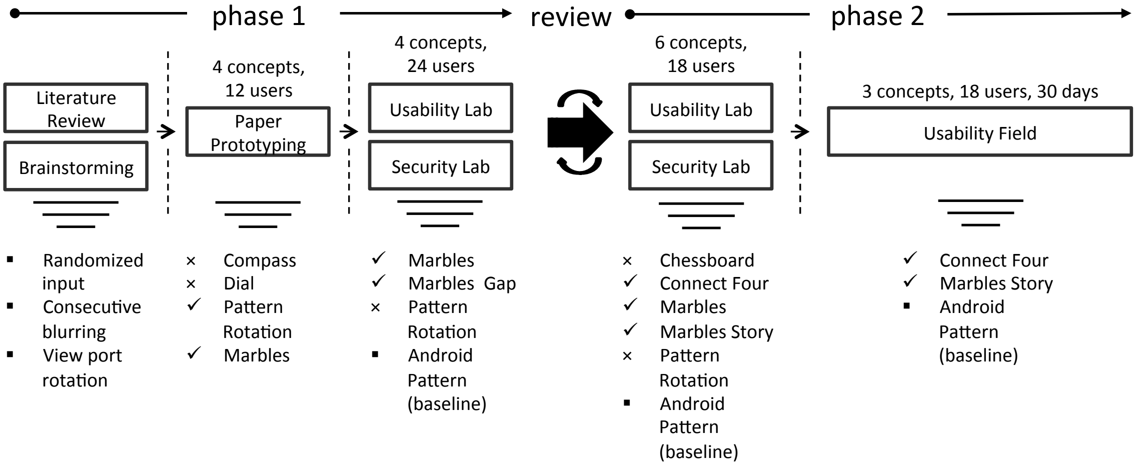


Figure 4.14: The concept development was divided into two phases and resulted in two promising authentication mechanisms: *Connect Four* and *Marbles Story*.

six-point Likert scales. Interactivity of the paper prototypes (e.g., rotation) was simulated by a researcher. All sessions were filmed for later analysis.

Participants: The concepts were tested by twelve experienced smartphone users. The average age was 22 (19-26) years. Seven participants were female, five were male.

Evaluation 2: First Lab Study

Based on the results of the paper prototyping study, four concepts were implemented as interactive prototypes. The prototypes were based on standalone applications which were optimized for the device used in the lab study (HTC Nexus One, Android OS v2.1).

Hypotheses: Five (main) hypotheses were defined for the lab study.

H1 The randomization of input elements increases smudge attack resilience

H2 The randomization of input elements reduces efficiency

H3 The randomization of input elements reduces effectiveness

H4 The specific type of randomization has no effect on usability

H5 The specific type of randomization has an effect on security

Design: The user study was based on a repeated measure factorial design. The independent variables were *system* with four levels (*Android unlock pattern* (baseline), *Pattern Rotation 90*, *Marbles*, *Marbles Gap*) and *password type* with two levels (*given, self-selected*). *System* was counterbalanced using a Latin square design, *password type* was alternated.

Procedure and Setup: Each session started with an *introduction* to smudge attacks. Each participant was assigned a unique ID which was used to specify the order of *system*. Each concept was tested twice (alternating *password type*) using the following procedure:

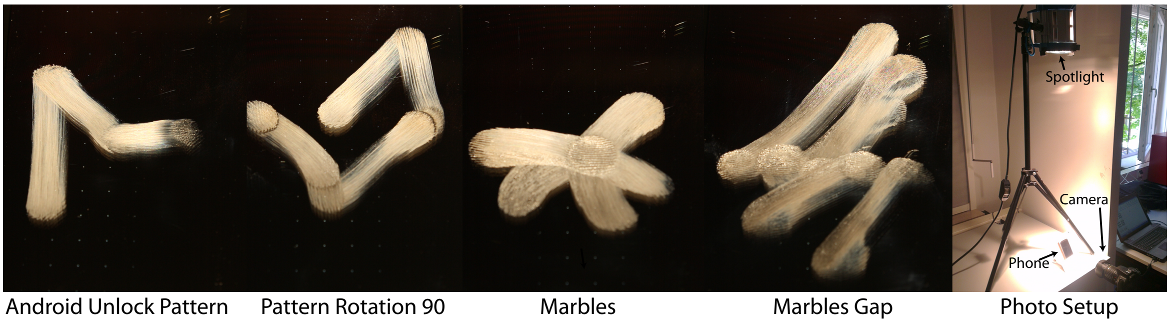


Figure 4.15: The Figure shows examples of the images used for the smudge attacks. The photo setup is shown on the right. The images have not been edited (except cropping).

Training The user tries out the respective system until she fully understands the approach.

Password Selection The user selects a password (or receives a predefined password) according to the policy stated below.

Cleaning The touchscreen is cleaned using a microfiber cloth.

Authentication The user enters the respective password. If the authentication fails, the process starts over with step 3: *Cleaning*.

Picture If the authentication was successful, a picture is taken.

Authentication The user has to successfully authenticate two more times.

During the first and the second authentication, users were allowed to look up their secrets. The third authentication had to be performed without memory aid. In the end of the study, we collected user feedback via questionnaire and participants were compensated with a 10 Euro shopping voucher. Based on the preferences observed in the preliminary study, passwords were restricted to the length of five. This resulted in a comparable theoretical password space for all concepts[4]. User interaction was filmed using a digital camcorder which was positioned behind the participant and targeted on the touchscreen of the device. As indicated by Figure 4.15, attacks were based on the approach presented by Aviv et al. [18]. We used a high-resolution camera (Canon EOS 1000D) and a 650W ARRI spotlight.

Participants: 24 participants were recruited via social networks and word-of-mouth advertising. The mean age was 25 (19-33) years. Eight users were female, 16 were male. All participants reported to be experienced touchscreen users. Thirteen (54%) subjects had already heard about smudge attacks. Seven participants stated to protect their smartphone with Android unlock patterns, six used PIN. The rest did not use secure lock mechanisms.

[4] Five activated cells result in 7,152 combinations for Android pattern and pattern rotation. Using five marbles results in 59,049 (15,120 without repeated colors) using the standard approach and 64,800 (30,240 without repeated colors) using *Marbles Gap*.

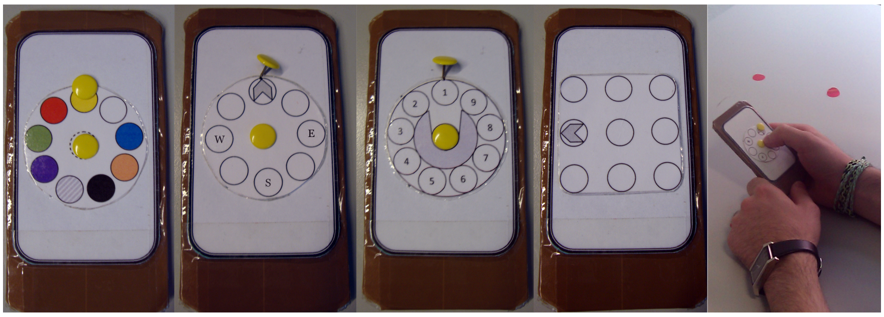


Figure 4.16: The Figure illustrates the paper prototypes were used in the first study. All prototypes provided flexible input elements which allowed to simulate authentication.

Usability Findings

The user studies provided various interesting insights into the relation between randomization and perceived efficiency. While this Section presents the most important results, more details can be found in [268].

Evaluation 1: Paper Prototyping
Figure 4.16 illustrates the paper prototypes used in the lab study. The results indicated that *Marbles* was the most promising concept. All participants acknowledged that *Marbles* was “very easy” to understand. Furthermore, all but one participant (who rated it neutral) attested “very good” usability. In addition, all but one participant stated that they would use *Marbles* on their personal devices. While participants found the rest of the concepts easy to understand as well, the overall rating was more negative. Even though *Dial* and *Compass* performed well in terms of usability, only eight participants would use *Compass* on a daily basis and only nine participants would use *Dial*. *Pattern rotation* was rated worst. Two participants explicitly disagreed that *Pattern Rotation 90* was usable, five participants disagreed that *Pattern 360* was usable. Users reported that both systems demand high mental load and consequently, only the minority could imagine using such concepts on their personal devices.

Evaluation 2: First Lab Study
Based on the promising results of the paper prototyping study, we decided to implement *Marbles*. During the review of the results, we came up with the idea of a different *Marbles* layout: *Marbles Gap*. Although we had no preliminary results according its usability, we decided to include *Marbles Gap* into the laboratory study. User interaction was very similar and thus we assumed comparable usability. Even though the preliminary results of *Pattern Rotation* were not promising, we opted to include the 90-degree version for two reasons: Firstly, the concept was very close to Android unlock gestures and thus allowed direct comparison. Secondly, we wanted to investigate the quantitative effects of viewport rotation and confirm the increased mental demand.

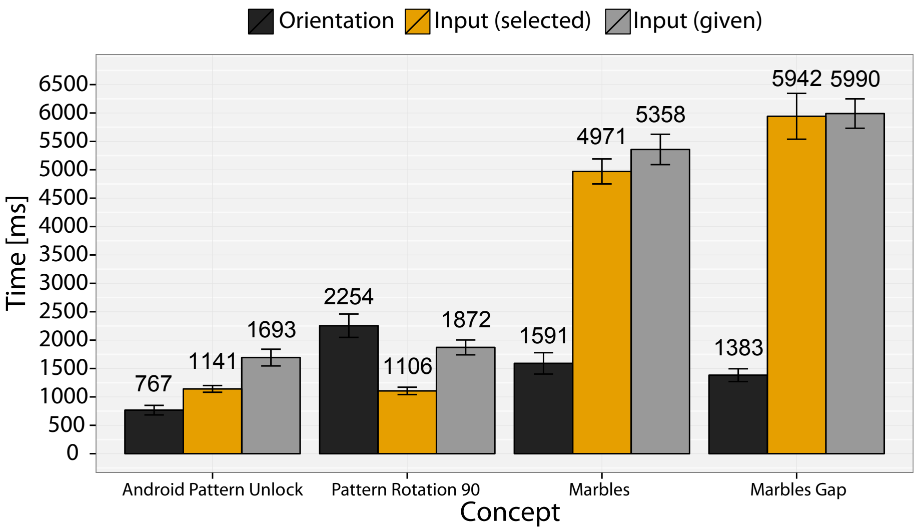


Figure 4.17: Efficiency was assessed based on authentication time. Splitting the authentication process into orientation phase and input phase provides valuable insights: Overall, the *Pattern Rotation 90* concept performs second best. However, orientation time exceeds input time.

Efficiency is assessed based on authentication speed. Authentication speed was measured as the sum of *orientation time* and *input time*. Orientation time represents the time a user needs to prepare the interaction (i.e., recognize the arrangement of the input elements). It was logged as the time span from the presentation of the authentication screen to the user's first touch event. The input time was measured from the first touch event and ended with the confirmation (or cancellation).

The results are based on 576 samples (*24 users* $*$ *2 password types* $*$ *4 systems* $*$ *3 authentications*). However, we filtered the failed attempts as well as extreme values that exceeded the doubled standard deviation as an upper or lower boundary. As mentioned, each test case was tested three times (3 authentications). Since a repeated measure ANOVA indicated no significant differences ($p > .05$) between those three steps, input time was analyzed based on the average of all three steps. According orientation time, we could only use the third authentication step and based the analysis on 192 samples (*24 users* $*$ *2 password types* $*$ *4 systems* $*$ *1 (last) authentication*)[5]. We performed a repeated measure ANOVA of the orientation times and input times. Figure 4.17 illustrates the results concerning the third run.

In contrast to *Password type* ($p > .05$), the used authentication *system* had a significant influence on the orientation periods, $F_{2.1,48.4} = 16.64, p < .001$, Greenhouse-Geisser corrected: $\varepsilon = .69$. Bonferroni-corrected post-hoc tests reveal that users needed significantly more preparation time whenever the spatial arrangement was randomized. *Pattern rotation 90* (Mn = 2254ms, SE = 20) required most time, followed by *Marbles* (Mn = 1592ms,

[5] The first and the second step were invalid due to logging errors. Nevertheless, we assume that the last run is well suited to assess efficiency as users were already familiar with the respective secret.

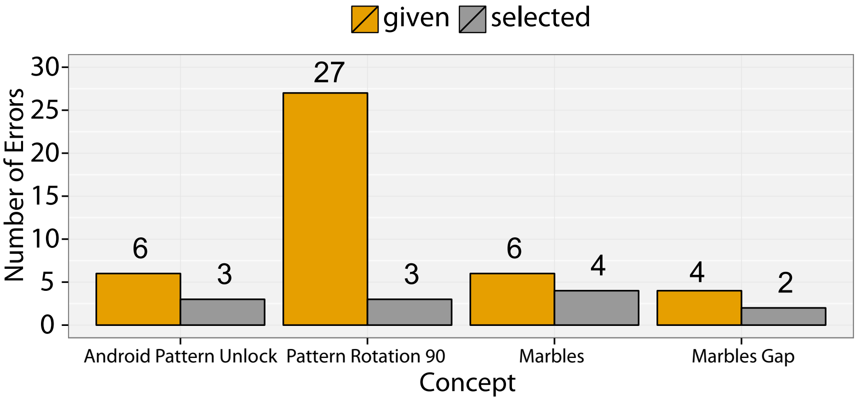


Figure 4.18: While the overall error-rate was 9.5%, *Pattern Rotation 90* was significantly more error-prone than any other system.

SE = 188) and *Marbles Gab* (Mn = 1383ms, SE = 113). Orientation times of *Android unlock patterns* were significantly shorter (Mn = 768ms, SE = 84) but *Marbles* and *Marbles Gap* performed significantly faster than *Pattern Rotation 90* ($p < .05$).

The input time was influenced by both independent variables as we found significant main effects for *system* ($F_{2.1,41.6} = 315.32, p < .001$, Greenhouse-Geisser corrected: $\varepsilon = .69$) and *password type* ($F_{1.0,20.0} = 27.25, p < .001$). The post-hoc analysis indicates that using *Marbles* (Mn = 5233ms, SE = 199) and *Marbles Gap* (Mn = 5982ms, SE = 261) demands significantly more time ($p < .001$) than *Android unlock pattern* ($Mn = 1611ms, SE = 111$) and *Pattern Rotation 90* (Mn = 1664ms, SE = 97). However, no significant differences were found when comparing marble-based and pattern-based approaches with each other ($p >$ .05). While marble-based approaches were not affected by *password type*, users performed significantly faster when self-selected gestures were used on grid-based concepts ($p < .001$).

Effectiveness was assessed based on the number of input errors (576 samples). Overall, we observed 55 failed attempts resulting in an error rate of 9.5%. In one instance, a participant was not able to authenticate at all. The user failed three times in a row using *Pattern Rotation 90* with a given password.

A repeated measure ANOVA analyzing the number of failed attempts indicates significant main effects for *system* ($F_{3.0,40.0} = 5.99, p < .05$, Greenhouse-Geisser corrected: $\varepsilon = .58$) and *password type* ($F_{1.0,23.0} = 8.15, p < .05$). The error rates are illustrated in Figure 4.18. Bonferroni-corrected post-hoc tests reveal that *Pattern Rotation 90* ($n = 30$) was significantly more error-prone than any of the other concept ($p < .05$ for all contrasts). The lowest error-rate was achieved by *Marbles Gap* ($n = 6$) while *Android unlock pattern* led to nine failed authentications and using *Marbles* resulted in ten input errors. Furthermore, given passwords led to significantly ($p < .05$) more errors as 78% ($n = 43$) were based on this *password type*. A detailed analysis of the errors reveals that 90% ($n = 27$) of the *Pattern Rotation 90*-errors were based on predefined gestures.

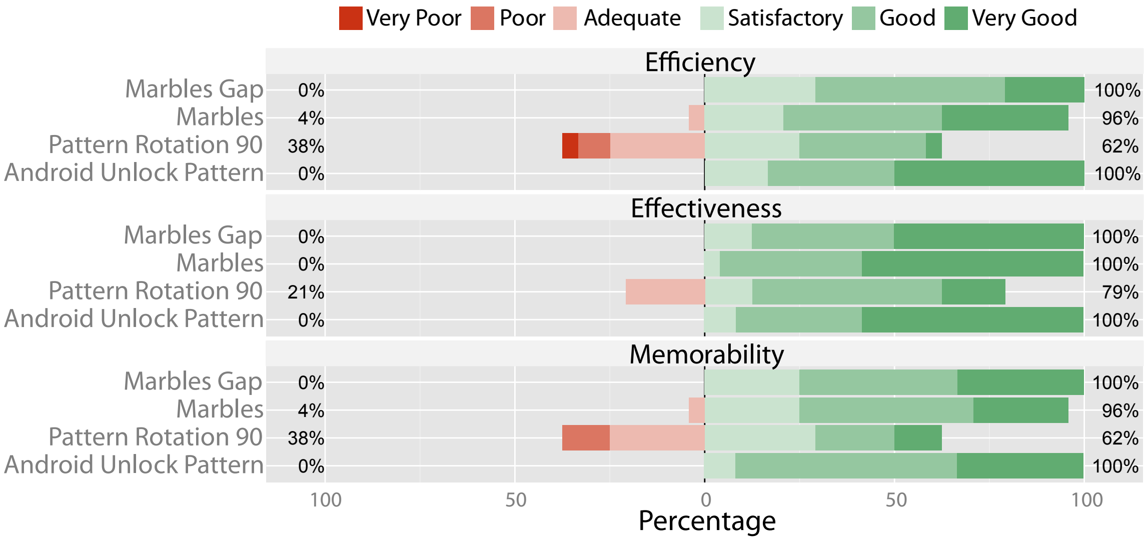


Figure 4.19: Perceived usability based on six-point scales. Users preferred the baseline as well as *Marbles*.

Perception and Likability were assessed based on 6-point scales and concept rankings. Figure 4.19 illustrates the results according to efficiency, effectiveness and memorability.

Interestingly, the perceived efficiency of the randomized concepts differs from the quantitative results. Even though *Pattern Rotation 90* performed second best in terms of measured authentication speed, it was rated worst by the users. The median of the ratings indicates "satisfactory" efficiency. Three users even disagreed that *Pattern Rotation 90* was efficient. In contrast, *Marbles* and *Marbles Gap*, which actually demanded most authentication time, were both rated "good". The static *Android unlock pattern* was rated consistently with the measured values. The individual rating was confirmed by the final rankings.

The ratings of effectiveness and ease of use are more consistent with the measured values. Users claimed that the rotation of the *Pattern Rotation 90* approach was "cumbersome" and "demanded high cognitive load". Participants were often confused and stated that most errors resulted from performing gestures in the wrong direction. The ratings illustrated in Figure 4.19 confirm these statements: *Pattern Rotation 90* was overall rated "satisfactory", while the other concepts were rated "good". Furthermore, participants confirmed that self-selected gestures were easier to perform: *Pattern Rotation 90* was rated "satisfactory" when used with self-selected gestures but "poor" otherwise. In the final ranking, *Marbles Gap* was voted best; followed by *Android unlock pattern*, *Marbles* and *Pattern Rotation 90*.

The interaction problems resulted in perceived memorability issues. Even if the type of secret is exactly the same for *Android unlock pattern* and *Pattern Rotation 90*, the memorability was perceived differently. Based on the median, *Android unlock patterns* were rated "very good" for self-selected gestures and "good" for given secrets. *Pattern Rotation 90* was rated "good" for self-selection but only "satisfactory" when gestures were assigned. Even though some participants informally mentioned that color-based secrets are odd, both marble-based approaches were rated “good” independently from *password type*.

In terms of likeability, most users were in favor of *Android unlock patterns* followed by *Marbles* and *Marbles Gap*. *Pattern rotation 90* was the least favored concept. Even though most (92%) of the users could imagine using *Android unlock patterns* on a daily basis, two participants would not use this concept as “it was not secure enough”. Furthermore, 75% of the users stated they would use *Marbles* and 67% could imagine using *Marbles Gap*. Criticism on marble-based concepts was mostly related to the use of color coded input elements. Participants, who did not want to use these concepts, suggested the use of numbers or symbols instead. Finally, 42% of the users could imagine using the *Pattern Rotation 90* concept on a daily basis. Participants, who did not want to use this approach, claimed that it required too much spatial imagination and rotating the mobile device itself was cumbersome.

Security Findings

Smudge-attack resistance was assessed based on the pictures taken during the user study. Figure 4.15 shows examples of the images used for the smudge attacks. The shown examples illustrate very clear smudge traces which are easy to identify. Nevertheless, the intensity of the smudge residues depends on the oiliness of the user’s fingers and therefore some images showed less clear traces. The security results are based on 192 samples (2 password types $*$ 4 systems $*$ 24 users) which were collected in the lab study. Attacks were simulated by a researcher who was familiar with the used concepts. The attacker was informed about the used *password type* but had no knowledge of the actual passwords. The attacker was allowed to zoom in and to rotate but no other transformations or adjustments were made. We allowed three guesses per image and calculated the success rate on a binary basis (true/false). In addition to the attacker, a second researcher was present to note down the results.

As assumed (*H1*) randomization of input elements increases smudge resistance. The static *Android unlock patterns* are very vulnerable as 83% of such gestures were exposed (independent from *password type*). The remaining 17% were resistant due to very dry fingers. The users did not leave enough oily residues for interpretation. The analysis of the number of guesses indicates that 60% of the gestures were identified after the first attack and the rest was found in in the second attack. The third attempt did not improve the success rate. This indicates that smudge attacks are successful whenever enough residues are left on the screen.

Pattern Rotation 90 increased the practical security as only 46% of the gestures could be derived. Assuming clear smudge traces, attackers had three attempts to guess one of four possible directions (rotations). This results in a theoretical chance of 75%. Indeed, the distribution of successful guesses indicates that the attacker randomly guessed the direction as 36% of the exposed gestures were found in the first guess, 32% were identified in the second attack and another 32% of the gestures were derived in the third guess. However, the attacker informally mentioned that in some cases smudge residues indicated how the device was grasped and that this cue could be used to derive the orientation of the matrix.

Both marble-based concepts performed very good and allowed no successful attacks. Nevertheless, it became apparent that attackers could use smudge attacks to derive meta information of the used secrets. This information can limit the search space for other attacks (e.g., educated guessing). For example, the limited reusability of *Marbles Gap* can expose sensitive details. If smudge traces stem from only one segment, the attacker knows that no color was used twice. The same is true for *Marbles*: As the spatial arrangement of the input elements remains fixed during one authentication, smudge traces can give useful cues on the diversity of the used tokens. In the worst case, if only one smudge trace is visible, an attacker could conclude that only one color was used to compose the secret. If the length were additionally known, the search space would be reduced to the number of input elements.

Summary and Implications

In this first development cycle, we designed several graphical gesture-based authentication concepts which utilize different kinds of randomized spatial arrangements. The concept candidates were implemented as low-fidelity paper prototypes and high-fidelity software prototypes. The prototypes were tested in two user studies to evaluate the usability and the security of the systems. The results indicated that the concepts are more secure against smudge attacks than Android unlock patterns (*H1*). While efficiency was indeed decreased (*H2*), we found interesting contrasts between the measured log data and the qualitative user feedback. While *Pattern Rotation* performed second best when efficiency was quantitatively assessed, it was rated worst. In contrast, *Marbles* was rated fast by the users, even if authentications took more than twice the time. In terms of effectiveness, we cannot clearly reject *H3* as the *Pattern Rotation* concept became very error-prone when it was used with assigned gestures. At the same time, all the other conditions performed as well as the baseline. We conclude that the type of randomization has a significant impact on usability (*H4*), especially on perceived performance. Finally, while all randomized concepts were more secure, the fully randomized *Marbles* concept worked significantly better (*H5*).

In addition, we gained valuable insights into the specific design of the concepts. For example, *Marbles* was criticized for its color-coded input elements. In the next development phase, we thus wanted to achieve mainly three goals: Firstly, *Marbles* needed to be further improved. Secondly, based on the lessons learned from the (failed) design of the *Pattern Rotation* concept, we aimed at developing an improved grid-based authentication system which was both more usable and more secure than *Pattern Rotation*. Finally, we desired to investigate the reasons for the observed contrasts between measured efficiency data and user perception.

4.2.5 Improving Smudge-resilient Gesture Input

While *Marbles* performed well in terms of security and usability, *Pattern Rotation 90* had serious drawbacks. The second development circle started with an ideation phase which built on the results of the first studies. As we were particularly interested in developing a feasible grid-based gesture concept, the ideation process resulted in two rotation-based schemes: *Chessboard* and *Connect Four*. In addition, *Marbles* was redesigned with concrete representations of input elements to improve both memorability and user experience. The concepts were again evaluated in a lab study before the most promising concepts were tested in the field. The first part of this Section presents the study design and procedure. The second part presents the results of the performed evaluations.

Prototypes and Evaluation Strategy

First, the concepts were implemented as interactive prototypes for Android and tested in the lab. *Android pattern unlock*, *Connect Four* and *Marbles Story* were then implemented as lock screen replacements and tested in a longitudinal field study. This Section provides details on the study designs. The results of both studies are represented in the next Section.

Evaluation 3: Second Lab Study

The second lab study was based on the approach of the first iteration.

Hypotheses: We defined the following main hypotheses for the second lab study.

H1-1 *Chessboard* and *Connect Four* are more secure than pattern rotation

H1-2 *Chessboard* and *Connect Four* are as usable as pattern rotation

H1-3 Using concrete representation of input elements improves the usability of *Marbles*

H1-4 The representation of input elements has no impact on smudge resistance

Design: The study was based on a repeated measure factorial design. The only independent variable was *system* with six levels (*Android unlock pattern* (baseline), *Pattern Rotation 90*, *Marbles*, *Marbles Story*, *Chessboard*, *Connect Four*). We opted to limit the evaluation to self-selected gestures as this approach was more consistent with the real world and equivalent to the later field study. However, selected gestures needed to correspond to a strict policy. *System* was counterbalanced using a Latin square design.

Procedure and Setup: The approach matched the first user study with two minor changes in the procedure. Firstly, the pictures for the later smudge attacks were taken after the performance test was finished. That is, after participants had authenticated three times, the device was cleaned and handed over for a fourth authentication. Secondly, the password policies were adjusted to satisfy a higher theoretical level of security. With the exception of *Connect Four*, gestures of the grid-based concepts needed to connect six cells. The two marble-based concepts presented ten input elements and users needed to select four items. *Connect Four* required the connection of five cells. The length was reduced as the additional uncertainty of the unknown anchor point increases the search space by the factor eight. Besides the new length requirements, composition policies remained the same. In addition to the modifications in the procedure, we used a slightly changed hardware setup. The concepts were tested using a HTC One (Android v. 4.1.2.) and pictures were taken with a Canon EOS 50D.

Instead of the Arri spotlight, a Bowens Gemini 500R flash light was used. Again, quantitative data was automatically logged and qualitative data was collected via questionnaires. Participants were compensated with a 10 Euro shopping voucher.

Participants: 18 participants were recruited via social networks and word-of-mouth advertising. The mean age was 26 years (23-32). Six users were female, twelve were male. All participants reported to be experienced touchscreen users. Five (28%) participants indicated to be concerned about smudge attacks. Ten (55.6%) users protected their device with a PIN-based lock screen and seven (38.9%) used Android unlock patterns. None of the participants had taken part in the user studies of the first iteration.

Evaluation 4: Longitudinal Field Study

The second lab study identified *Marbles Story* as the most promising concept. In addition, we opted to evaluate *Connect Four* in the field. It was the only concept which provided both smudge attack resistance and observation resistance. Furthermore, we desired to analyze if the usability of such concepts could be improved if they were used over a longer period of time. The main goal of the field study was to investigate learning effects and to assess the suitability for daily use.

To increase the external validity of the study, we replaced the lock mechanism of the users' devices. Therefore, participants used the tested concepts with their personal smartphones and integrate them into their daily routines. As a consequence, the prototype needed to support different devices with different form factors. The application was developed for Android v. 2.3 and higher. After the application was registered as device administrator, it was able to replace the original lock screen. The prototype was tested in a one week pre-study. In addition, we implemented a client-server architecture to remotely collect logging data and to control important study factors (e.g., which concept was used). The communication was based on an SSL-encrypted connection.

Hypotheses: We defined the following hypotheses for the field study.

H2-1 All concepts allow that authentications become more efficient over time.
H2-2 All concepts allow that authentications become more effective over time.
H2-3 *Connect Four* allows better learning effects than *Marbles Story*.
H2-4 Learning effects result in higher user acceptance rates.

Design: The user study was based on a repeated measure longitudinal design. The only independent variable was *system* with three levels (*Android unlock pattern* (baseline), *Marbles Story*, *Connect Four*). *System* was counterbalanced using all possible permutations. The user study lasted 30 days while each concept was tested for ten days.

Procedure and Setup: Participants were recruited based on a questionnaire. All participants had to own an Android smartphone with Android v. 2.3 or higher. After selection, participants were invited to the lab for an individual meeting. Within this briefing, we introduced the concepts, explained the study procedure and configured the participant's device. When the application was opened for the first time, users were asked to select a PIN or password as fallback secret. This secret was used whenever the authentication with the primary concept failed (e.g., due to memorability problems). Next, the first concept was presented and participants were asked to select a primary secret. To increase comparability, the secret had to comply with the same policy as used in the lab study. The concept was then tested for the next ten days. After ten days, participants answered a questionnaire. Whenever the questionnaire was completed, we remotely activated the next concept and users were again asked to select a secret. After all three concepts were tested, users were invited for a personal debriefing. Within the debriefing, the application was uninstalled and participants answered a final questionnaire. In the end, participants were compensated with a 20 Euro shopping voucher. During the 30 days of the user study, we logged primary authentications, fallback authentications and configuration changes (e.g., setting a new secret). However, we did not log the actual passwords.

Participants: We started with 19 participants and finished with 18 valid data sets. The average age was 27 (23-32) years. Six participants were female, twelve were male. All participants owned an Android smartphone and used it on a daily basis. Most (67%) participants used Android unlock patterns to protect their device. None of the participants had taken part in one of the prior user studies.

Usability Findings

This Section presents the usability results of both the second lab study and the field study.

Evaluation 3: Second Lab Study

The data is analyzed in terms of efficiency, effectiveness and user perception. Each concept was tested 54 times (18 participants $*$ three authentications).

Efficiency was assessed analyzing the user-based average of three successful authentications. Authentication speed was again split into orientation time and input time. The analysis is based on a repeated measure ANOVA.

The authentication *system* had a significant impact on orientation time, $F_{2.8,48.5} = 28.54, p < .001$, Greenhouse-Geisser corrected: $\varepsilon = .57$. Bonferroni-corrected post-hoc tests reveal that *Android unlock patterns* (Mn = 906ms, SE = 124) performed significantly faster than all other concepts ($p < .05$). *Marbles* (Mn = 1565ms, SE = 130) and *Marbles Story* (Mn = 1574ms, SE = 123) performed second best. Users needed significantly less time with marble-based concepts than with rotation-based concepts ($p < .05$). *Pattern Rotation 90* (Mn = 2974ms, SE = 333), *Chessboard* (Mn = 3226ms, SE = 234) and *Connect four* (Mn = 3653ms, SE = 320) demanded most orientation effort. Notably, no significant differences were shown between*Pattern Rotation 90*, *Chessboard* and *Connect Four* ($p > .05$).

Furthermore, the analysis revealed a significant main effect on input times, $F_{1.7,29.4} = 9.29, p < .05$, Greenhouse-Geisser corrected: $\varepsilon = .35$. Bonferroni-corrected post-hoc tests showed that *Android unlock patterns* (Mn = 1293ms, SE = 88), *Connect Four* (Mn = 1350ms, SE = 115) and *Pattern Rotation 90* (Mn = 1588ms, SE = 164) performed

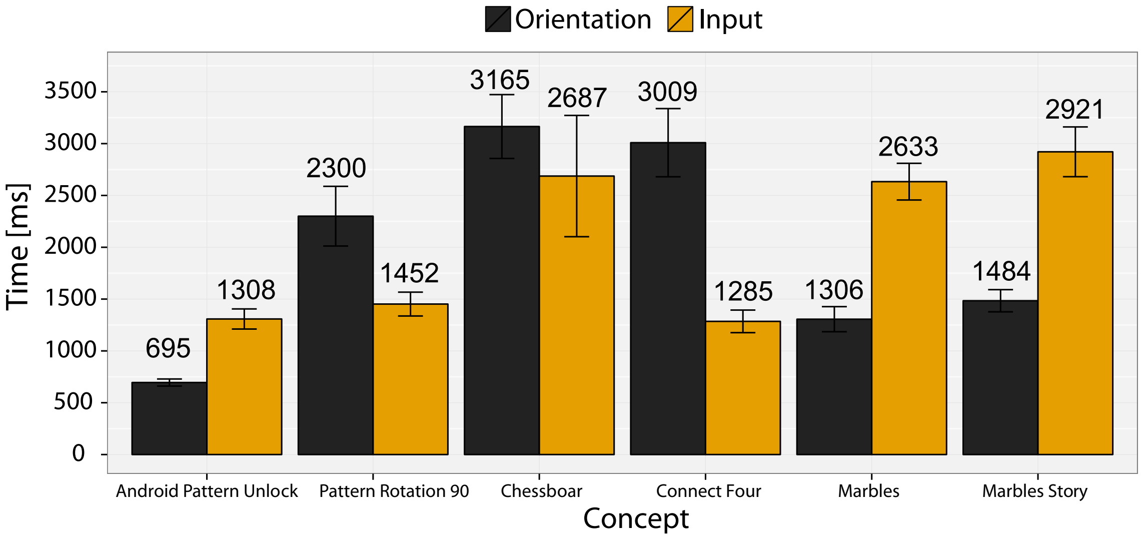


Figure 4.20: The *Android pattern unlock* performs best in terms of orientation times and input times. Marble-based approaches have advantages considering the orientation phase, rotation-based approaches tend to perform well in terms of input times.

significantly faster than marble-based approaches ($p < .05$). Even though the mean input times indicated that *Chessboard* (Mn = 2834ms, SE = 579) was similarly time-demanding as *Marbles* (Mn = 2838ms, SE = 209) and *Marbles Story* (Mn = 3080ms, SE = 233), no significant differences were found when it was compared to the other rotation-based concepts ($p > .05$). Figure 4.20 illustrates the measured times of the third authentication. However, due to the different password length, the comparability to the results of the first lab study is limited.

Effectiveness was assessed based on the number of errors. Overall, we observed 57 failed attempts resulting in an error rate of 17.6%. Nine (17%) inputs failed with *Android unlock pattern*. Twelve (22%) errors were logged with *Pattern Rotation 90*, 15 (28%) authentications failed when *Connect Four* was used and 23 (43%) errors occurred while using *Chessboard*. Both marble-based approaches were less error-prone: Five (9%) errors were logged with *Marbles*, two (4%) authentications failed with *Marbles Story*.

Perception and Likability was assessed using Likert scales and the final concept ranking. Figure 4.21 illustrates the user feedback according to efficiency, effectiveness, memorability and likeability. The overall feedback is analyzed based on the median values. *Android unlock pattern* and *Marbles Story* score best in terms of performance as most users fully agreed that the concepts are efficient, effective and memorable. In addition, most users fully agreed that using *Marbles Story* was fun. The abstract *Marbles* approach was also rated good but participants confirmed (slight) memorability drawbacks. The randomized grid-based approaches were rated worse in terms of efficiency and effectiveness. This particularly confirms the findings of the first lab evaluation that rotation-based concepts are perceived

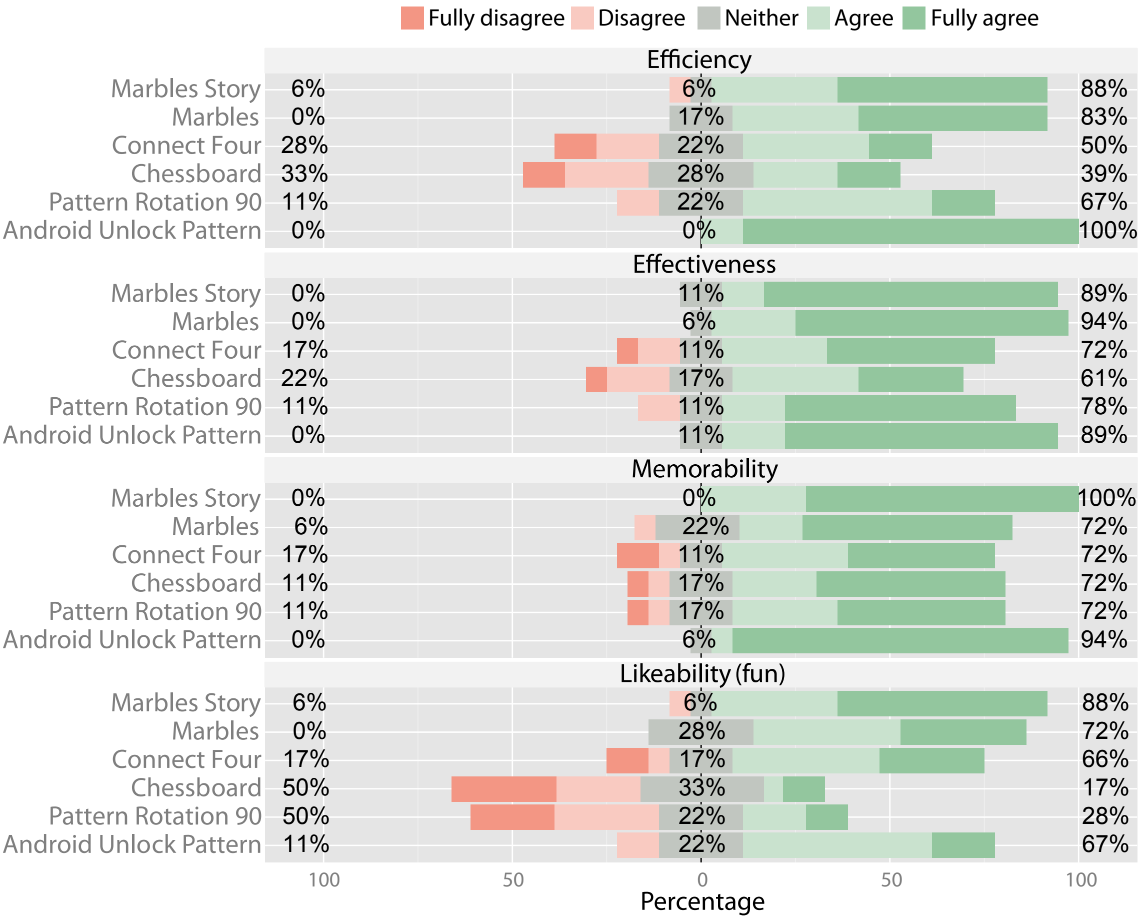


Figure 4.21: Perceived usability ratings based on five-point Likert scales. While *Marbles Story* scored best, the randomized grid-based concepts were perceived inefficient and ineffective. Still, *Connect Four* could outperform *Pattern Rotation 90* and *Chessboard* in terms of likeability.

disproportionately slow. Nevertheless, *Connect Four* scored in terms of likeability as most users agreed or strongly agreed that using this concept was fun.

Finally, ten users acknowledged that they would use *Android unlock patterns* on a daily basis. Even more participants stated the same for *Marbles* and *Marbles Story* as eleven users would use *Marbles* and 16 users would install *Marbles Story* on their devices. The rotation-based concepts scored worse. Even if using *Connect Four* was fun, only three participants would use it on a daily basis. The same was true for *Pattern Rotation 90* and *Chessboard*.

Evaluation 4: Longitudinal Field Study

Based on the results of the second lab study, we finally selected *Marbles Story* and *Connect Four* for the longitudinal field study. The goal of the study was to analyze if people can get used to the randomized layouts and how orientation times and input times are affected by training effects. Over the course of 30 days, we logged 22,061 authentications. Partici-

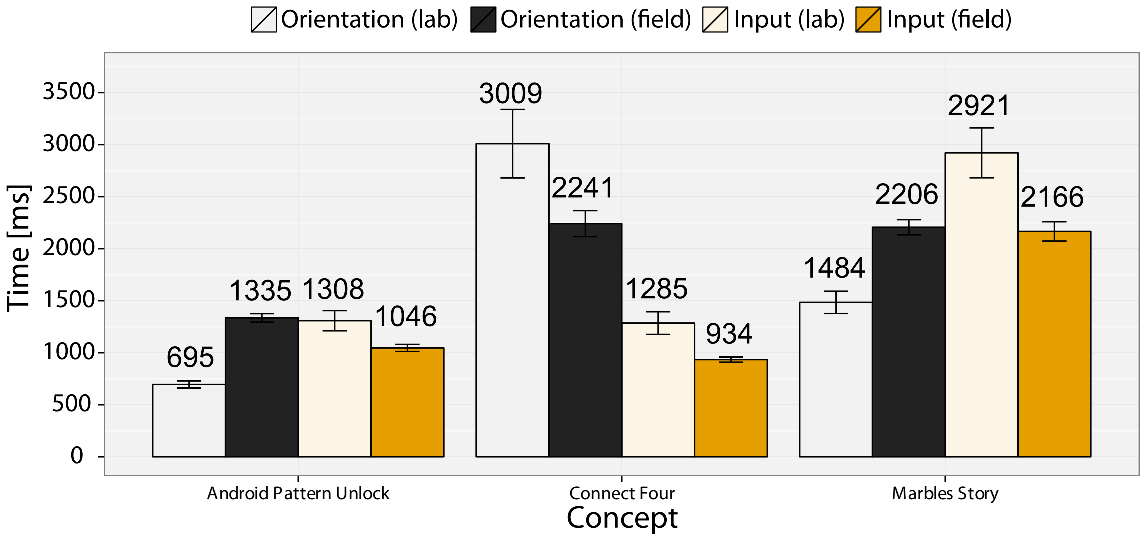


Figure 4.22: The Figure illustrates the mean authentication time of the second lab experiment and the field study. While input times generally decreased in the field, the approximated orientation times decreased only for *Connect Four*.

pants performed 7,419 unlocks using *Android unlock patterns*, 9,578 unlocks using *Connect Four* and 5,064 authentications using *Marbles Story*. For each user, we included the first seven days per concept which comprised a minimum of two authentications. The data was averaged per user and per day.

Efficiency was assessed based on the average authentication time of successful authentications. Again, authentication time was split into orientation and input time. Orientation time started with screen-on events and ended with the first touch event. As indicated by Section 3.2, mobile devices are often turned on to check notifications. As a consequence, not every screen-on event is immediately followed by an authentication and measured orientation times may include other actions than preparation. To approximate orientation effort, we excluded outliers by setting a cut off. The considered maximum was set to the doubled maximum of the orientation time observed in the second lab study.

Figure 4.22 illustrates the average authentication speed of the field study and the second lab study. A repeated measure ANOVA considering the average daily orientation times of each user reveals a significant main effect for the authentication system, $F_{1.4,21.2} = 40.07, p <$ .001, Greenhouse-Geisser corrected: $\varepsilon = .71$. Post-hoc tests reveal that users needed significantly less orientation time using *Android unlock patterns* (Mn = 1335ms, SE = 41) than using the other two concepts $p < .001$. However, no significant differences were found between *Marbles Story* (Mn = 2207ms, SE = 73) and *Connect Four* (Mn = 2242ms, SE = 125), $p > .05$. In the second lab study, *Marbles Story* (Mn = 1574ms, SE = 123) had performed significantly faster than *Connect Four* (Mn = 3653ms, SE = 320). Figure 4.23 illustrates the measured average on a daily basis. Independently from the used concept, orientation effort

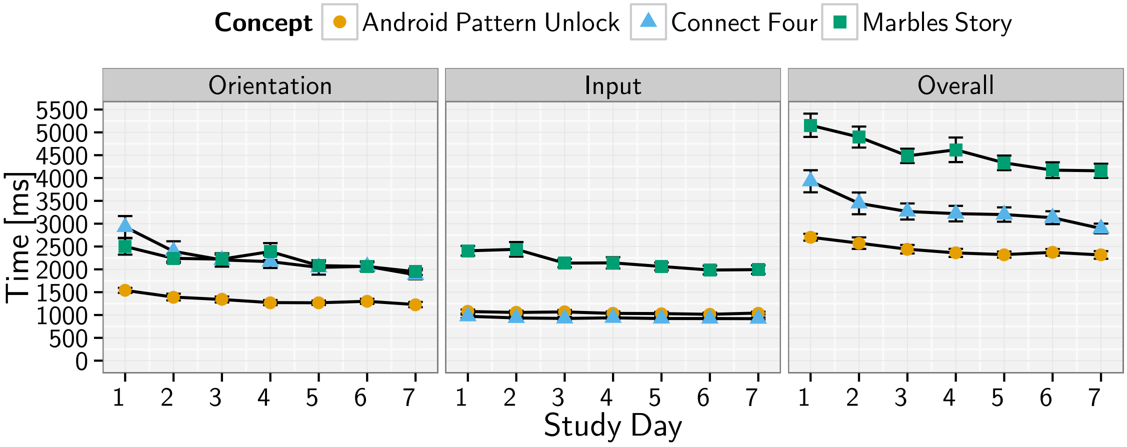


Figure 4.23: Over the course of seven days, orientation times tend to decrease independently from the used system. Furthermore, the input times of grid-based concepts remain constant while the marble-based concept allowed faster input over time.

decreased over time. However, we found no significant interaction effect between *system* and *day*, $F_{4.0,60.3} = 2.43, p > .05$, Greenhouse-Geisser corrected: $\varepsilon = .34$.

According to input effort, all three concepts performed faster in the field than in the lab study. A repeated measure ANOVA comparing the average input times of each user per day reveals a significant main effect of the used authentication *system*, $F_{1.1,15.8} = 121.93, p < .001$, Greenhouse-Geisser corrected: $\varepsilon = .57$. Bonferroni-corrected post-hoc tests reveal that input times using *Connect Four* (Mn = 934ms, SE = 25) were significantly shorter than input times using *Android unlock patterns* (Mn = 1046ms, SE = 34) and *Marbles Story* (Mn = 2166ms, SE = 93), $p < .05$. Furthermore, users performed significantly faster using *Android unlock patterns* than using *Marbles Story*, $p < .05$. In addition, we found significant interaction effects of *days* and *system*, $F_{3.9,54.9} = 5.94, p < .05$, Greenhouse-Geisser corrected: $\varepsilon = .33$. As indicated by Figure 4.23, users became faster over time using *Marbles Story*.

A final ANOVA comparing the total authentication times of all concepts revealed a significant main effect of *system*, $F_{2.0,30.0} = 74.88, p < .001$. Corrected post-hoc tests indicate that *Android unlock patterns* (Mn = 2441ms, SE = 65) allow the fastest authentication. Followed by *Connect Four* (Mn = 3298ms, SE = 141) and *Marbles Story* (Mn = 4545ms, SE = 157) . Overall, authentication times differed significantly between all three systems.

Effectiveness was assessed on the averaged portion of successful authentications. As every participant performed a different number of authentications each day, the absolute number of errors had limited value. A repeated measure ANOVA revealed a significant main effects of *system*, $F_{1.3,20.1} = 31.94, p < .001$, Greenhouse-Geisser corrected: $\varepsilon = .63$. The average success rate using *Android unlock patterns* was 82% ($Ci_{95} = 77 - 87$). The success rate of *Connect Four* was 72% ($Ci_{95} = 63 - 81$) and therefore significantly lower, $p < .05$. *Marbles*

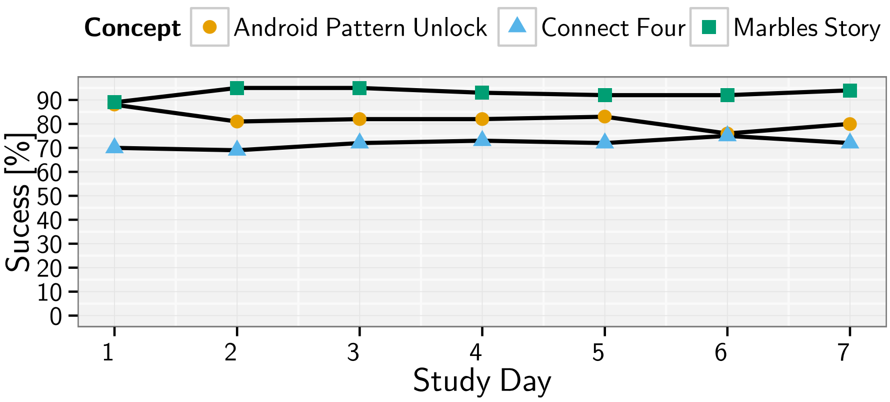


Figure 4.24: The success rate differs significantly between concepts. However, success rates remain rather constant over time.

Story was most effective with an average success rate of 93% ($Ci_{95} = 89 - 96$). Figure 4.24 illustrates the average success rate for each system over the course of the seven days. While the success rate differs between the tested concepts, it remains rather constant over time.

Perception and Likability was assessed using five-point Likert scales ranging from "fully disagree" to "fully agree". Figure 4.25 illustrates the user feedback according to efficiency, effectiveness, memorability, likeability, learnability and acceptance. Even though *Marbles Story* performed significantly worse than *Connect Four*, most users would agree that the system is efficient. In contrast, effectiveness was rated more consistent with the measured data. While *Marbles Story* was rated best, *Connect Four* was rated more negative as 39% of the users would "disagree" or "fully disagree" that the system was easy to use. Memorability was rated good for all systems. Interestingly, most participants favored the novel concepts. While 50% would "disagree" that using *Android unlock patterns* was fun, 72% liked using *Connect Four* and 78% agreed that *Marbles Story* was fun. To assess learnability, we asked participants if authentications became easier and faster over time. While most participants rated the performance constant when using *Android unlock patterns*, slightly more users would agree that learning effects kicked in when using *Connect Four* or *Marbles Story*.

Finally, we asked participants if they would use the respective concept on a daily basis. While 39% (7) of the participants "agreed" or "fully agreed" that they would use *Android unlock patterns* on their personal devices, 39% (10) stated the same for *Connect Four* and *Marbles Story*. At the same time, a substantial portion of the users disagreed concerning all three concepts. This indicates that user acceptance is not generalizable and hard to predict.

Security Findings

Smudge attack resilience was analyzed following the approach of the first iteration. The pictures of the second lab study were analyzed and attacked by a researcher who was highly

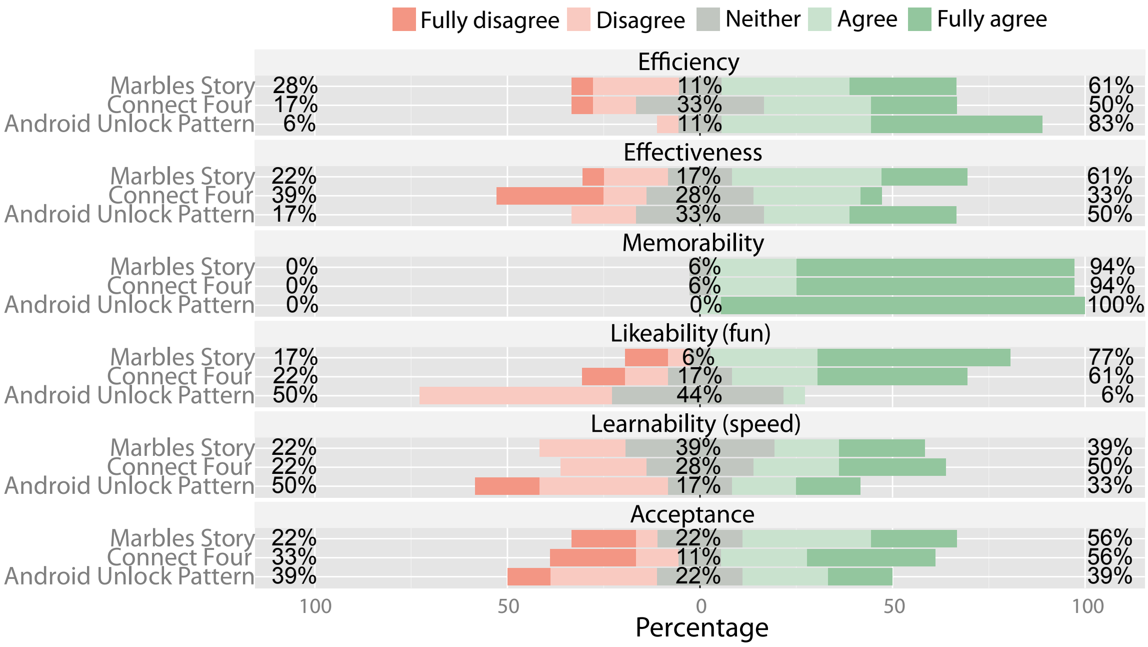


Figure 4.25: While the perceived effectiveness matched the measured data, efficiency was rated in favor of *Marbles*. In addition, the novel concepts scored in terms of likeability.

familiar with the concepts but had no knowledge of the used secrets. As the first security evaluation had shown that smudge traces of the marble-based approach reveal no information on the used secret, *Marbles* was excluded from the experiment. The chance of finding the right combinations would be identical to an uninformed guessing attack.

The rest of the concepts were attacked after performing the lab study. The attacker was able to identify 100% (18) of the secrets entered with *Android unlock patterns*. *Pattern Rotation 90* revealed 72% (13) of the entered gestures. In contrast, modifications of *Pattern Rotation 90* performed better. Nevertheless, 28% (5) of the gestures could be derived when *Chessboard* was used. Focusing only on the attacked concepts, *Connect Four* performed best as only one (6%) instance was successfully attacked. Even though the security of *Marbles Story* was not empirically assessed, chances are high that none of the entered secrets would have been guessed.

Summary

Based on the results of the first development cycle, we started with a thorough revision of the concepts. We designed two modifications of the *Pattern Rotation* concept and improved the visual representation of the *Marbles* concept. All concepts were implemented as software prototypes and evaluated in the lab and in the field. While the second lab study provided further valuable insights into the impact of specific design decisions, the field study shed light on learning effects and real world performance.

The second security analysis confirmed that both rotation-based concepts provide improved smudge attack resilience (*H1-1*). However, the lab evaluation also indicated a decrease of usability as both concepts demanded more authentication time (compared to *Pattern Rotation 90*) and users failed more often. Therefore, hypotheses *H1-2* cannot be accepted. Nevertheless, users were in favor of *Connect Four* and stated that using the concept was fun. According to *Marbles Story*, the results were clearer. The lab study confirmed that *Marbles* benefits from a concrete representation of input elements as it increased memorability and user acceptance improved (*H1-3*).

Due to the promising results, especially in terms of user experience and security, we decided to test *Connect Four* and *Marbles Story* in a longitudinal field study. Indeed, it was shown that both randomized concepts allow authentications to become more efficient over time (*H2-1*). Interestingly, the fully randomized *Marbles* concept allowed faster input while *Connect Four* seemed to support shorter orientation phases. In addition, the existence of training effects was indicated by the fact that all three concepts performed faster in the field than in the lab. While the data revealed that *Marbles Story* allows higher success rates than the static *Android unlock patterns*, effectiveness remained constant over time. Therefore, *H2-2* cannot be accepted. Similarly, we cannot accept the hypotheses that *Connect Four* supports stronger learning effects than *Marbles Story* (*H2-3*). Finally, the real-world application showed that the new systems are widely accepted. Indeed people liked the novel systems more than the static *Android unlock patterns* (*H2-4*).

4.2.6 Discussion and Implications

In this Section, we summarize the main findings and discuss their implications.

Randomized Spatial Arrangements Can Increase Smudge Attack Resistance

The findings confirmed that *Android unlock patterns* are vulnerable to smudge attacks and thus confirm the findings of Aviv et al. [18]. In the first security evaluation, we were able to identify 83% of the entered gestures, in the second analysis all gestures were found. The novel randomized authentication mechanisms were significantly more secure against smudge attacks. At the same time, we found that the type of randomization has a significant impact on security. While rotation-based approaches like *Pattern Rotation 90* are still prone to smudge attacks, the smudge traces of marble-based concepts did not reveal any information. This indicates that the randomization of input elements is preferable to the randomization of the view port. In addition, *Connect Four* indicated that the temporal arrangement of the challenges is another important factor. Technically speaking, *Connect Four* is based on the randomized spatial arrangement of input elements. However, gestures are still entered in one *single* challenge. Since this aspect would still allow effective smudge attacks, secret guidance was implemented as additional security factor. We conclude that *randomized spatial arrangements* are most effective when combined with *multiple* input challenges.

Randomized Spatial Arrangements Can be Efficient and Effective

As assumed, the evaluation showed that *randomized spatial arrangements* tend to downgrade the efficiency of the authentication. Overall, authentications using randomized methods took more time than authentications using *Android unlock patterns*. However, a detailed analysis of authentication times indicated that the concepts influenced efficiency in different ways. Using rotation-based authentication methods (e.g., *Pattern Rotation 90*) resulted in increased orientation times, while marble-based concepts tended to increase input times. The phenomenon can be explained by the different *temporal arrangements*. The single-challenge procedure of rotation-based concepts requires more preparation effort. However, as soon as the current setup is understood, gesture input is as efficient as with static concepts. In contrast, marble-based projects split the preparation effort over multiple challenges. With each challenge, the current marble is located and logged. In all studies, rotation-based concepts were significantly more efficient than marble-based concepts. In the longitudinal field study, users needed about 3.3 seconds to authenticate using *Connect Four* and 4.5 seconds using *Marbles Story*. We therefore conclude that both concepts are in acceptable range (cf. Section 3.2) and that randomized spatial arrangements indeed allow efficient authentication.

In terms of effectiveness, we conclude that the overall error rate was low. However, we observed one exception in lab study one where the high error rate of *Pattern Rotation 90* indicated low effectiveness. A detailed analysis revealed that 90% of the errors occurred while using predefined gestures. We assume that participants did not correctly memorize such predefined gestures and consequently mixed up directions more often. Marble-based approaches indicated high effectiveness and even outperformed the *Android pattern unlock*. We therefore conclude that randomized spatial arrangements have no negative impact on effectiveness when self-selected or trained gestures are used but rotation-based systems tend to be harder to use.

Temporal Arrangement Influences User Perception and Performance

Despite being measurably more efficient, rotation-based approaches were constantly rated slower than *Marbles*. Even after one week of use, *Marbles Story* was perceived more efficient than *Connect Four* although the latter allowed significantly faster authentications. A detailed analysis of the authentication times shows that user ratings rather focus on the *orientation times* than on the whole authentication effort. Indeed, all rotation-based authentication methods which were based on *single* input challenges, demanded significantly higher orientation effort before the first input could start. As a consequence, *orientation times* sometimes even exceeded *input times*. In contrast, the *temporal arrangement* of the *Marble* concepts led to a better relation between *orientation* and *input times* as using *Marbles*, the first input could start quicker and the overall orientation time was split into multiple shorter phases (one for each challenge). The effects were confirmed in the second lab study and in the field.

The results indicate that high orientation times are more annoying for users than high input times and have a significant impact on the overall rating of efficiency. Neurobiological experiments have indicated that time spans are perceived as longer if “more contextual changes are available for retrieval" [34,293]. Such context changes can be triggered by "different cognitive load variables (e.g., , degree of task difficulty)" [34]. In addition, it was found that expectations and interruptions stretch the perceived temporal length of a given task [101]. Using rotation-based concepts, both aspects seem fulfilled: Due to the single-challenge setup, there is a significant contextual change between the orientation task and the input task. Furthermore, the perceived time seems to be additionally stretched as users expect to start with the gesture. Using *Marbles*, the context remains constant and the expected interaction task can start quicker as orientation effort is split into multiple challenges. We conclude that generally, the orientation effort should be minimized. Especially, orientation effort should not exceed input effort. Moreover, we recommend to design the whole authentication task in a way which minimizes contextual changes. The evaluation of marble-based concepts indicated that randomized spatial arrangements are perceived faster, when orientation tasks can be distributed over *multiple* challenges.

Spatial Arrangement and Representation can Influence Memorability and Learnability

It is often assumed that memorability is mainly affected by the type of secret. However, the evaluation revealed that *Pattern Rotation 90* and *Android unlock patterns* were differently perceived in terms of memorability. Despite using the same grid-based gestures, users rated the memorability of *Pattern Rotation 90* worse. While the spatial arrangement during gesture selection is identical, the difference is found during authentication. The randomized spatial arrangement of *Pattern Rotation 90* results in four different gestures (based on the same shape). As a consequence, gesture recall becomes more difficult and quick learning may be hindered. At the same time, we assume that the user will learn four distinct gestures in the long run and memorability problems will be reduced. Interestingly, memorability of *Marbles* was rated very well even though these concepts do not support motor memory. In addition, user acceptance was improved by *Marbles Story* which provided concrete representations of input elements and thus allowed story-based recall of secrets.

We conclude that memorability problems may arise if the spatial arrangements of the enrollment and the authentication are different. In addition, randomized spatial arrangements have a negative impact if gestures are based on shapes (e.g., grid-based). Nevertheless, rotation-based concepts may support motor-memory effects and therefore allow more efficient authentications in the long run.

Randomization can Decouple Input Complexity and Password Strength

With regard to gesture selection, randomized authentication concepts provide several interesting aspects. The results indicate that self-selected gestures are more efficient and more effective when used on grid-based concepts. In contrast, such effects were not observed when *Marbles* was used. We assume that this phenomenon is based on the fact that grid-based gestures provide a wider range of input complexity compared to marble-based passwords. The fully randomized spatial arrangement of *Marbles* decouples input complexity and password strength. That is, the gesture is not influenced by the complexity of the chosen secret.

We conclude that fully randomized spatial arrangements lead to a more counterbalanced use of secrets while the selection of grid-based gestures is biased due to varying input complexity. However, marble-based concepts might have a negative impact on gesture length as the use of additional input elements linearly increases search times and input effort.

Longitudinal Field Studies and Accurate Data Collection are Crucial

User studies are vital and lab evaluations are important to give first insights into the usability and the security of novel concepts. However, authentication concepts are built for real life and should therefore be tested in real settings. While a field study does usually not allow fine-grained analyses of efficiency and effectiveness, the ecological validity gives important insights into the real world suitability of concepts. We observed that randomized spatial arrangements allow training effects and that both final concepts were feasible in the wild. User acceptance and efficiency actually increased compared to the lab study and one participant kept using *Connect Four* after the end of the study. According to security, we like to note that lab evaluations allowed to simulate a worst case scenario. As a consequence, it was justified to assume that no attacker would perform better in the wild and the field evaluation could focus on usability aspects. We conclude that, whenever possible, authentication systems should be evaluated in the wild. Security evaluation can form an exception, if worst case scenarios (attacker's best case) can be simulated in the lab.

In addition, the evaluation showed the importance of accurate measurements and confirmed that usability should be assessed quantitatively and qualitatively. In particular, we found interesting contrasts between measured and perceived efficiency values. Due to the consideration of all stages of the authentication process, we were able to interpret such contrasts. Therefore, we conclude that orientation and clean up phases must be considered.

Lessons Learned from the Rejected Concepts

We argue that it is important to report and discuss the whole design process as crucial lessons are learned from rejected concepts. Even though the results indicated that *Pattern Rotation 90* is neither usable nor secure, the evaluation of the concept provided important insights into the interplay of spatial arrangements and temporal arrangements. The same is true for the other rejected concepts. For example, *Chessboard* taught us that not every additional randomization aspect does necessarily increase security. *Marbles Gap* indicated an important interplay of input effort and password selection which can inspire future concepts. Finally, the report of failed ideas might prevent other researchers from making the same mistakes. We therefore conclude that the discussion of rejected concepts is almost of the same value as the discussion of promising concepts.

4.2.7 Limitations

Even though the user studies were carefully designed and conducted, the approach had some inherent limitations which will be addressed here.

We had to accept some restrictions concerning the external validity to increase the internal validity and to collect comparable data. Most notably, gesture selection was guided by strict policies. While such composition rules allow better comparison between concepts and between setups (e.g., lab versus field), they prevent insights into the users' selection behavior.

According to the field study, it is crucial to keep in mind that the quantitative assessment of performance can only be based on approximated values. As indicated by Chapter 3, mobile interaction is usually a secondary task and often interrupted. As a consequence, we assume that participants did not always authenticate as fast as possible. However, we argue that such interruptions occur independently from the used concept and therefore the observed relations are not affected by such errors. In addition, the static grid-based approach performed very similar to the approach presented in Section 3.3. Since the user study presented in Section 3.3 was based on a more controlled field design, we conclude that the data is valid.

Concerning the presented concepts, we do not claim that we found the perfect solutions. We rather illustrated feasible approaches and explored the design space. Even though, the systematic approach resulted in two very promising concepts, there are probably other concepts which would work equally well. Finally, we have to acknowledge that all user studies were based on a limited set of self-recruited users. Overall, participants were younger and more tech-savvy than it would be expected from the general population. We therefore argue that the results can give general insights on the usability and the security of randomized concepts but the results are not directly transferable to other populations.

4.2.8 Summary

In this Section, we explored to design space of graphical gesture-based authentication mechanisms with the aim of increasing smudge resistance. Utilizing an iterative design process, we confirmed the vulnerability of *Android unlock patterns* and proposed several alternative concepts. The proposed solutions were implemented as low-fidelity paper prototypes and high-fidelity software prototypes and then evaluated in the lab and in the field. In addition to the presentation of two promising authentication mechanisms which were shown to meet the requirements of mobile interaction, the Section contributed by providing general insights into the design and evaluation of *randomized authentication methods*.

The evaluation revealed that *randomized spatial arrangements* can indeed be utilized to build smudge attack-resilient but yet usable authentication mechanism for mobile devices (*RQ1*). Based on the literature review and initial brainstorming, we proposed three general approaches to achieve smudge attack resistance: *Consecutive blurring*, *viewport rotation* and *randomized input elements*. Fully randomized arrangements of input elements (e.g., *Marbles*) worked best when multiple challenges were considered. In addition, we showed that rotation-based single-challenge concepts (e.g., *Connect Four*) can be usable and secure when specific design conditions are fulfilled (e.g., unchanged context, additional cues).

Comparing quantitative and qualitative data, we found important aspects concerning perception and user acceptance (*RQ2*). Since users do not want to turn the device before input, rotated view ports were perceived more cumbersome than (multiple) sequential search tasks. The analysis indicated that perceived efficiency was mainly influenced by two factors: *orientation effort* and *context changes*. High orientation times seem more annoying than high input times and a changing context between orientation and input tasks seems to further stretch experienced time. Since user perception is a critical factor for user acceptance, both aspects need to be optimized. That is, orientation times need to be reduced and the context should remain stable. In addition, the results showed that user acceptance can be further increased by using colorful and playful representations of input elements.

While randomization generally increased authentication times, effectiveness was not affected (*RQ3*). The field study revealed that error rates were very low for all tested concepts and *Marbles* did even outperform *Android unlock patterns*. Nevertheless, the results of the first user study indicated that rotation-based concepts are significantly more error-prone when more complex gestures are assigned. Future work needs to investigate if users indeed opt for easy-to-enter gestures when view port rotation is applied as this might reduce the practical gesture space. In contrast, the *Marbles* concept illustrated that fully randomized arrangements can decouple input complexity and gesture selection.

Finally, the field evaluation confirmed that randomized authentication concepts support learning effects and become more efficient over time (*RQ4*). While memorability was not systematically evaluated, user feedback indicated that all concepts were memorable and easy to learn. Furthermore, the design of *Marbles Story* illustrated that concrete representations of input elements can further improve such aspects.

4.3 On Preventing Observation Attacks

This Section explores the design space concerning *observation attacks*. Section 3.4 revealed that currently used gesture-based authentication mechanisms are very much prone to such attacks and Chapter 2 already introduced alternative concepts which were developed to make gesture input more resistant. However, the discussion indicated that most concepts are too cumbersome and not efficient enough to meet the requirements of daily mobile interaction. In this Section, we present a novel class of observation resistant authentication mechanisms. The proposed concepts allow to adjust the usability-security trade-off to the current context and remain very efficient in most situations.

We shed light on the following main *research questions*:

RQ1 How can we utilize gestures to allow *adjustable* authentication mechanisms on mobile devices which remain very efficient in most situations?

RQ2 How can we utilize *dynamic guidance* to build efficient and observation-resistant authentication methods?

RQ3 How can we design *self-contained* gestures that allow eyes-free and observation-resistant authentication?

RQ4 How do *directness* and *relation* influence observation-resistance and usability?

The design space exploration resulted in two concepts: *XSide* and *SwiPIN*. *XSide* prevents observation attacks by shifting parts of the authentication to the back of the device. Therefore, attackers need to observe both sides of the mobile device to succeed. We designed a novel class of gestures, built two prototypes and evaluated the concept in two *lab studies* (n = 56). *SwiPIN* allows observation-resistant PIN-entry by utilizing dynamically guided gesture input. While the interaction is too complex for observers to follow, users are not required to memorize any additional information. *SwiPIN* was developed in multiple iterations and evaluated in four *lab studies* (n = 56) and one *field study* (n = 12).

This Section is partly based on three research papers: 1) De Luca, A., von Zezschwitz, E., Nguyen, N. D. H., Maurer, M. E., Rubegni, E., Scipioni, M. P., & Langheinrich, M. (2013, April). Back-of-device authentication on smartphones. In Proceedings of the CHI'13. [74]. 2) De Luca, A., Harbach, M., von Zezschwitz, E., Maurer, M. E., Slawik, B. E., Hussmann, H., & Smith, M. (2014, April). Now you see me, now you don't: protecting smartphone authentication from shoulder surfers. In Proceedings of CHI'14 [71]. 3) von Zezschwitz, E., De Luca, A., Brunkow, B., & Hussmann, H. (2015, April). SwiPIN: Fast and secure pin-entry on smartphones. In Proceedings of CHI'15 [262]. In addition, parts are based on a bachelor thesis by Annika Busch [48] and a practical research project by Miriam Mickisch [182] which were both carried out under my constant supervision. Please refer to the beginning of this thesis for a detailed statement of collaboration.

In contrast to Section 4.2, all presented concepts are based on *fixed* input elements and utilize *multiple* challenges. We will learn how *adjustable* authentication mechanisms can increase usability whenever additional security is dispensable but improve observation-resistance when needed (*RQ1*). *SwiPIN* will illustrate that *dynamic guidance* can make gesture-input secure and usable when needed (*RQ2*). In addition, the design of *XSide* will show how *self-contained* gestures need to be designed to allow eyes-free interaction (*RQ3*). This new gesture-class allows observation-resistant authentication without the need of randomization. The discussion of both concepts will illustrate the impact of specific design decisions concerning *directness* and *relation*. For example, we will learn that users tend to use direct gestures whenever possible and that this aspect can significantly compromise security (*RQ4*).

4.3.1 Research Context and Motivation

Since Chapter 2 already covers related work in detail, this Section gives a short summary and focuses on the novelty of the herein presented concepts.

The discussion of observation-resistant authentication mechanisms (see Section 2.3) indicated that such concepts often introduce time-consuming input tasks (e.g., [207]), additional secrets (e.g., [72, 258]) or require elaborated hardware (e.g., [281]). As a result, security-optimized concepts are often perceived as cumbersome, inefficient and error-prone. On the other hand, Section 3.2 revealed that serious shoulder surfing risks are perceived rather seldom and that users often authenticate in trusted environments. This indicates that mobile authentication methods need to be very efficient to get the chance of a wide user acceptance and improved security is usually no selling point. Based on such results, the concepts presented in this Section were especially designed to be effective, efficient and easy-to-deploy.

XSide utilizes relative gestures which can be entered either on the back or on the front of the device. Therefore, *XSide* allows the user to establish a "secret channel" by hiding the input and can be compared to concepts like CoverPad [288]. Outside of the security context, back-of-device interaction has been proposed to reduce occlusion problems [22, 280] or to improve one-handed interaction with large devices [166]. Evaluating back-of-device interaction, Hasan et al. [119] found that relative gestures are more efficient than absolute input. Even though elementary interaction can already be enabled on most off-the-shelf devices [284], more elaborated devices are becoming available: For example, the YotaPhone[6] provides an additional e-ink display on the back of the device which supports input and output. While mobile device interaction can generally benefit from allowing input on both sides of the device [289], *XSide* was the first authentication concept to exploit such features.

SwiPIN is based on "visual overload and distraction" and can be compared to concepts like ColorSnakes [112]. While the normal interaction is based on common PIN-entry, the secure mode utilizes relative goal-oriented gestures. The expected gestures are communicated to the user via dynamic guiding elements. Similar to TinyLock [161], *SwiPIN* shows that

[6] `https://yotaphone.com` – accessed: 2016/04/11.

slight modifications of the user interface can significantly improve observation resistance. In contrast to previous work, *SwiPIN* does not require multiplexed input (e.g., [207]), additional secrets (e.g., [72]) or multi-modal interaction (e.g., [30]). Observation-resistance is achieved solely by switching from direct taps to relative gestures. The concept was featured on the technology blog "Gizmodo" as one of "12 Fascinating Projects From the Bleeding Edge of Interaction Design"[7].

In summary, we illustrate how relative gestures can be used to provide adjustable, efficient and effective observation-resistance on mobile devices. In addition, we show how gestures can be used to support eyes-free interaction and how gesture can be utilized to authenticate with traditional concepts (e.g., PIN). In contrast to the previous Section, the concepts presented in this Section comprise only little randomization and support a recall based on motor or visual memory.

4.3.2 Threat Model

According to the threat model, the observer is in the direct vicinity to the victim. One possible scenario could be based on an attacker observing mobile device users on public transport. The attacker can take any position (standing, sitting) and has perfect sight on the victim's device as there are no occlusions or reflections. As indicated by Section 3.2, malicious attacks are seldom. Therefore, we assume a casual observer performing a cognitive attack. The input is only observed once and no technical equipment is involved. After the authentication was observed, the attacker gains possession of the victim's device and tries to authenticate.

We assume that in such a (semi-)public scenario, the victim would be aware of potential observers and would thus opt for the more secure but less efficient interaction method. While at home or in other trusted environments, the user would opt for the more efficient but less secure interaction style.

4.3.3 Concept Overview

Both concepts provide ways to adjust the level of security to the current needs. As a consequence, authentication can be performed very fast in most situations and slower (more secure) interaction is only used if justified. Table 4.2 categorizes the concepts.

XSide: Back-of-Device Gestures

XSide is presented in Section 4.3.4. The system achieves observation resistance by utilizing a "secret channel". This secret channel is established through eyes-free interaction: Using *XSide*, gestures can be entered on both sides of the device. In risk-free situations, users

[7] `http://gizmodo.com/12-fascinating-projects-from-the-bleeding-edge-of-inter-1700656949` – accessed:2016/04/11.

		XSide	SwiPIN
Input	*Representation*	abstract	abstract
Elements	*Reusability*	unlimited	unlimited
Output	*Guidance*	none	dynamic
Elements	*Feedback*	aggregated	aggregated
Interaction	*Relation*	self-contained	target-oriented
Style	*Directness**	(in)direct	indirect
Element	*Temporal*	multiple	multiple
Arrangement	*Spatial*	fixed	fixed

Table 4.2: Both concepts are based on multiple challenges which are entered using a fixed spatial layout. **XSide* supports direct and indirect input. Even though preliminary instances of *SwiPIN* allowed direct gestures, the final concept was based on indirect input.

can perform gestures on the front screen (which is assumed to be efficient and error-free). However, if observation risk is perceived, users can adjust the level of security by using the back of the device. A common scenario could be a user sitting in the bus, as gestures are entered out-of-sight (on the back of the device), the secret is not exposed to the people next to the user. If required, input can even be distributed using the front and the back of the device.

Figure 4.26 illustrates the concept. To authenticate, a user enters n gestures (i.e.: $n = 3$). Gestures represent arbitrary combinations of horizontal and vertical strokes. Each gesture can be entered either on the front or on the back of the device. Since gestures are *self-contained* and described by relative direction changes, they allow effective input out-of-sight as neither the size nor the position of the entered gesture is relevant. To increase observation resistance, the concept forgoes *visual representations* of input elements. Input elements are described as interactive areas on the front and on the back of the device, input is confirmed using *aggregated feedback*.

Figure 4.27 illustrates the gesture concept (called *BoD-Gestures*) in more detail. The available alphabet consists of four directions: **U**p, **R**ight, **D**own, **L**eft. While directions can be combined to individual gestures, the actual secret is based on multiple instances of such single gestures. For example, the secret used in Figure 4.26 is based on three individual gestures and would be described as **U**,**RU**,**R**. As *BoD-Gestures* are translated to strings, the concept supports securely encrypted storage. In the user studies, we tested secrets based on three gestures with up to three direction changes each. Assuming this configuration, *BoD-Gestures* provide a theoretical password space of $52^3 = 140,608$ secrets[8].

[8] As direction changes matter, combinations like *UU* are not supported. As a consequence, there are 4 single-stroke gestures, $4 * 3$ double-stroke gestures, and $4 * 3 * 3$ triple-stroke gestures. That is 52 gestures overall.

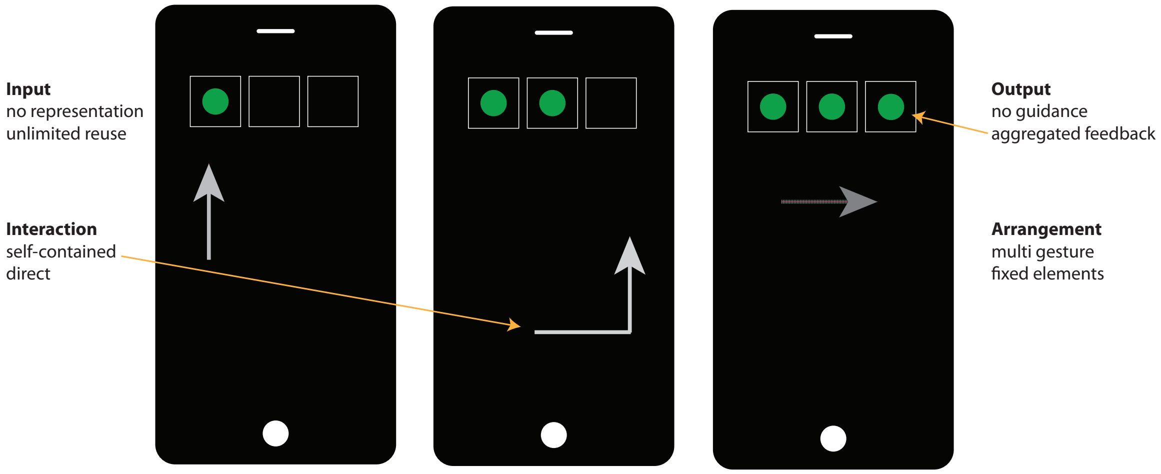


Figure 4.26: *XSide* is based on self-contained gestures which can be entered on the both sides of the device. Input elements are fixed but not visualized. The image on the right represents a gesture which was entered on the back of the device.

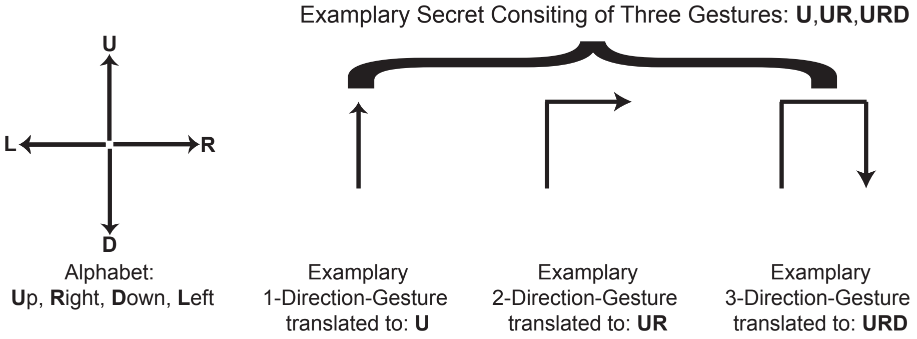


Figure 4.27: Single *XSide*-gestures comprise an arbitrary number of vertical and horizontal strokes. Every direction change counts as single stroke and is translated to a character (U,R,L,D). Secrets are based on multiple instances of individual gestures.

SwiPIN: Secure Gesture-based PIN-Entry

SwiPIN is presented in Section 4.3.5. The concept is based on "visual overload". In trusted environments, the system provides a common PIN-pad and users can efficiently authenticate by directly entering their PIN. Whenever observation risks are present, users switch to the more secure mode where digits are entered using relative gestures. *SwiPIN* utilizes a small redundant set of five simple gestures: *up*, *right*, *down*, *left* and *tap*. As the assignment of gestures and digits changes after each input, the entered PIN is very hard to observe.

Figure 4.28 illustrates three different layouts of *SwiPIN*. To map ten digits to five gestures, the *SwiPIN*-pad is subdivided into two differently colored sections. Each Section presents five digits, mapped to a distinct set of gestures. The current assignment is indicated by *dynamic guiding* elements which are represented by black arrows. Whenever no arrow is

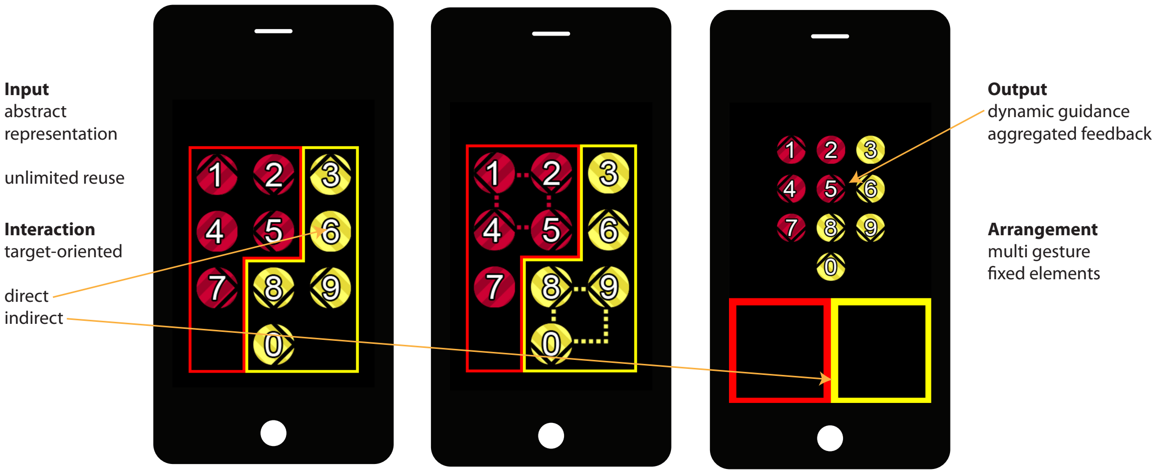


Figure 4.28: The *SwiPIN*-pad is subdivided into two differently colored sections. Each Section presents five digits, mapped to a distinct set of five gestures. The assignment is indicated by dynamic guiding elements which are represented by black arrows.

present, the "tap" gesture was assigned. Input elements are fixed and represented by colored boarders. To authenticate, the user has to (a) recognize the current mapping and (b) perform the assigned gesture. The gesture has to start inside of the input element but can end anywhere on screen. The assignments change after each input.

Let us now assume, a user wants to enter the digit "5" using the layout on the right. She would focus the respective digit and (a) recognize the arrow directed to the right. As the digit five is part of the red Section, the user would (b) perform a gesture to the right starting in the red input area on the bottom of the screen. As soon as the input area is touched, the mapping (guidance) disappears. The digit "5" is logged in and the system is ready for the next input. To allow efficient orientation, the color mapping, the order of the digits and the input elements are fixed. However, the gesture assignment changes after each input. That is, after "5" was entered, the mapping would change and performing the same gesture in the same input area would probably trigger a different digit.

The specific combination of *dynamic guidance* and *static input elements* allows users to efficiently authenticate. In addition, there is no need to memorize any additional information. In contrast, observers would need to memorize the assignments to be able to interpret the entered gesture. We assume that this task is usually too complex for cognitive observers. As illustrated in Figure 4.28, the three layouts utilize different input areas. The concept on the left (called *SwiPIN free*) allows to start gestures anywhere in the respective Section of the PIN pad. *SwiPIN inside* requires the user to start the gestures within the input elements represented by dotted borders. The layout on the right (*SwiPIN outside*) isolates input and output areas. As illustrated by the example above, users start their gestures within the colored squares at the bottom of the screen. The evaluation will show that such *indirect* gesture input can improve both security and user acceptance.

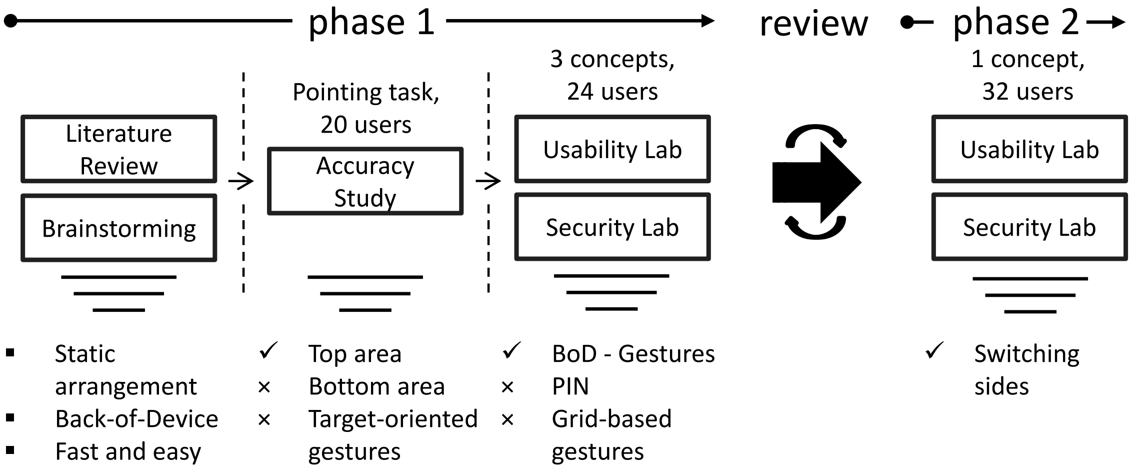


Figure 4.29: Concept development was based on two phases: In the first development cycle, we developed the gesture concept and compared it to common authentication methods. In the second phase, we evaluated the effects of side switching using an improved physical prototype.

4.3.4 XSide: Designing Gestures for Back-of-Device Authentication

As illustrated by Figure 4.29, concept development of *XSide* was divided into two phases. The first development cycle started with a literature review and ideation sessions. The brainstorming was based on the prerequisite that the interaction style allows efficient and effective back-of-device input and resulted in the *BoD-Gestures* presented in Section 4.3.3. After building a basic hardware prototype, we performed a preliminary accuracy study. Next, we performed a fist lab study to compare *XSide* to PIN and grid-based gestures. In the second development cycle, the hardware prototype was improved and a second lab evaluation was performed to assess the impact of side switching. We firstly present the prototypes and outline the evaluation strategies. The results of the performed studies are discussed in the second part of this Section.

Prototyping and Evaluation Strategy

XSide was developed based on an initial brainstorming and three laboratory user studies. This Section describes hypotheses as well as the evaluation strategies. Since *XSide* demanded sophisticated hardware prototyping, we start with a description of the hardware.

Prototypes and Implementation

One of the main challenges of the project was based on the fact that no feasible device was available for purchase[9]. As a consequence, we needed to build a device which allowed back-of-device interaction. We opted to simulate an interactive back side using a second smartphone. Over the course of the project, two prototypes were built (see Figure 4.30).

[9] As indicated in Section 4.3.1, we assume that this will change within the next years.

First Prototype: The first version of the prototype was rather basic. We glued two hardcover protective shells to each other and inserted two HTC One S smartphones. The smartphones were mounted back-to-back and rotated by 180° to equalize the bulges of the camera lenses. Hardware buttons and the lower 60% of the back side were covered with rubber band to allow proper grasping and to prevent unintended activation. The prototype, which is illustrated in Figure 4.30, was 1.5 cm thick and weighed 269 grams.

Second Prototype: Based on the lessons learned from the first two user studies, we built an advanced prototype. The prototype matched the look and feel of a real smartphone to improve user experience and to increase usability. We used 3D printed cases and thinner smartphones. We opted for the Alcatel One Touch Idol Ultra which was advertised as the "the slimmest smartphone in the world". Weight was further reduced by removing the back covers and the camera lenses. Finally, the interactive back side was increased to 50%. The second prototype was 1.2 cm thick and weighed 247 grams (see Figure 4.30).

Software: The communication between the two devices was established using Wi-Fi Direct. While both devices ran the same application, roles (front, back) where automatically assigned on start up. Touch events of the rear device were then sent to the front device and translated into the local coordinate system. Computation was handled by the front device: *BoD-Gestures* were extracted and interpreted using the ShortStraw algorithm [282]. The recognized strokes were matched to one of four directions and undefined combinations were deleted (e.g., "up,up"). In the second iteration, gesture recognition (and error resistance) was further improved by excluding all strokes that were shorter than 25% of the average stroke length or shorter than 60 pixels in total.

Evaluation 1: Accuracy Study

The accuracy study was performed to inform the position of the input elements and to assess the feasibility of accurate pointing and dragging on the back of the device.

Design: The study was based on a repeated measure within participants design. The independent variables were *target position* with eight levels, *dragging direction* with two levels and *grasping style* with two levels (*one-handed*, *freestyle*). Each dragging task (*target position* $*$ *dragging direction)* was performed once for each *grasping style*. While the input tasks were randomized, the *grasping style* was counterbalanced.

Procedure and Setup: The display of the back-of-device prototype was subdivided into $4x2 = 8$ evenly sized squares, each representing one level of *target position*. The back side was not covered by rubber band. Every task displayed two targets (circles) (labeled: "1" and "2") on the front screen. Participants pointed at target "1" and dragged it to target "2" using the back of the device. Failed tasks were repeated up to three times. After a training phase, 224 dragging operations were performed, half of them using only one hand. Whenever participants lifted the finger, the outcome was indicated changing the target color. However, the system did not provide any detailed feedback (e.g., virtual pointer).

Participants: 20 participants were recruited using mailing lists. The average age was 26 years (19-38). Thirteen were male, seven female.

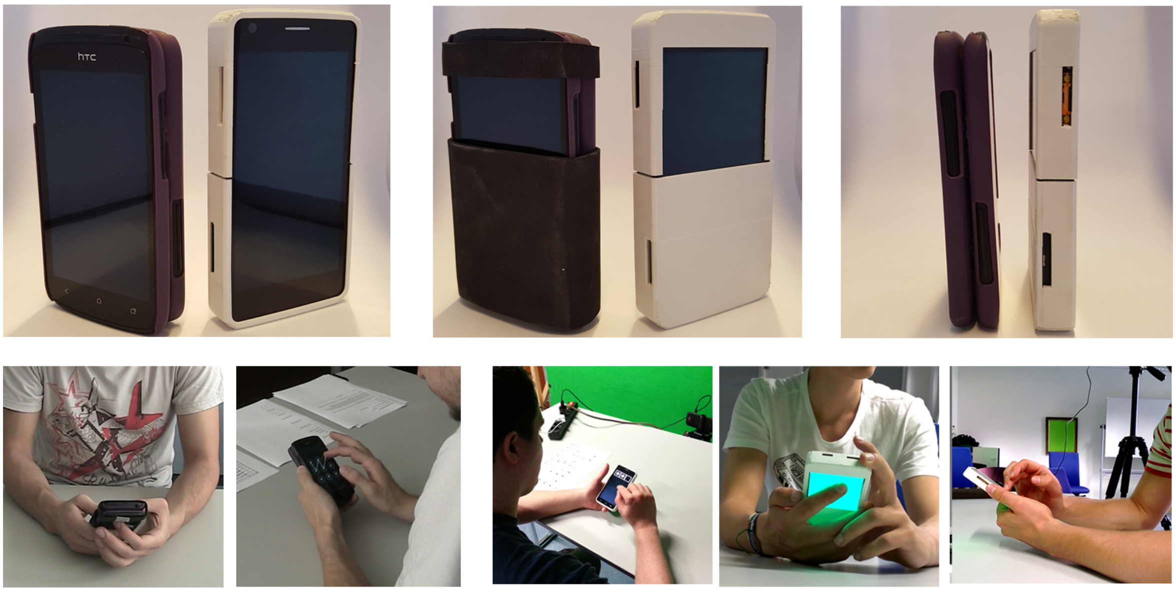


Figure 4.30: The first row shows two *XSide* prototypes with the white prototype being the improved version. The lower row illustrates the camera setup used for the observation attacks. The two images on the left originate from the first user study, the three pictures on the right show participants of the second study.

2. Evaluation: First Lab Study - Concept Comparison

The goal of the first lab study was to prove the feasibility of *BoD-Gestures* and to compare them to regular PIN-entry and to grid-based gestures. In addition, we tested a front-side version, called *Front-Gestures* to isolate the effects of both the gesture concept and the back-of-device interaction.

***Hypotheses**:* Five (main) hypotheses were defined for the first lab study.

H1-1 *BoD-Gestures* allow effective and efficient authentication on the back of the device.

H1-2 Back-of-device interaction increases observation resistance.

H1-3 Back-of-device interaction reduces the efficiency and effectiveness.

H1-4 *PIN* and *Grid-unlock* are more efficient and more effective than *Front-Gestures* and *BoD-Gestures*.

H1-5 *Front-Gestures* are as vulnerable to observation attacks as *PIN* and *Grid-unlock*.

***Design**:* The study was based on a repeated measure factorial design. The independent variables were *system* with four levels (*PIN*, *grid unlock*, *BoD-Gestures*, *Front-Gestures*), *secret type* with two levels (*given*, *self-selected*) and *secret complexity* with two levels (*easy*, *hard*). *System* was counterbalanced. *Secret type* and *secret complexity* was randomized.

PINs were based on four digits, grid-gestures used six cells and BoD-Gestures (and Front-Gestures) were based on three challenges. *Easy* secrets were composed in a way that reduced input effort. Easy *Front-Gestures* and *BoD-Gestures* comprised two single-stroke gestures and one two-direction gesture. Hard instances were built upon one three-direction gesture and two two-direction gestures. Hard PINs included only different digits and hard grid-gestures included a knight move. Composition rules applied to *given* and *self-selected* secrets.

Procedure and Setup: The user study was held in an isolated room at our premises. The *XSide*-Prototype was equipped with the four authentication systems. PINs, grid-based gestures and *Front-Gestures* were entered using the front of the device, *BoD-Gestures* were entered on the back side. User interaction was logged for later analysis.

The session started with an introduction of the study goals and a presentation of the concept. Afterwards, the first system was explained in detail and the participant performed a training task. Whenever the participant felt ready, the study began. Every concept was tested with four secrets (*secret type* ∗ *secret complexity*). Secrets were entered three times, failed authentications were repeated with a maximum of three attempts. After all concepts were tested, user feedback was collected via questionnaire and participants were compensated with a 5 Euro shopping voucher.

The videos for the security analysis were recorded with two cameras (see Figure 4.30, left). One camera simulated an attacker from behind, the other camera simulated an attacker sitting across from the user. The camera positions represented best case scenarios for the attacks with perfect sight on the interaction areas. We did not note the security analysis but mentioned that the recording was part of the usability analysis.

Participants: We recruited 24 participants via mailing lists and word-to-mouth advertisement. The average age was 27 (21-33) years. Eight participants were female, 16 were male. All participants used touch-based devices on a daily basis and were familiar with touch screen interaction for several years (Mn = 3; SD = 1.7). Eighteen (75%) participants used secure lock screens, nine of them used gesture-based concepts.

3. Evaluation: Second Lab Study - Side Switching

The first lab evaluation revealed that back of device authentication is usable and secure. The second study aimed at investigating the effects of side switching. We assumed that using both sides within one authentication attempt might further increase observation resistance.

Hypotheses: Two (main) hypotheses were defined for the second lab study.

H2-1 Higher numbers of *switches* reduce efficiency and effectiveness.

H2-2 Higher numbers of *switches* increase observation resistance.

Design: The user study was based on a repeated measure factorial design. The independent variables were *begin* with two levels (*front*, *back*) and *switches* with four levels (*0*, *1start*, *1end*, *2*). *Begin* specifies the side of the first gesture input, *switches* defines the number and position of switches. As all secrets were based on three gestures, the independent variables covered all possible combinations. For example, *back*∗*1end* required the user to start on the back of the device and switch before the last gesture is entered. Combinations with *switches = 0* represented the baselines: *front only* and *back only*. The independent variables were counterbalanced using an 8 × 8 Latin square design. All secrets were predefined and comprised two-direction gestures.

Procedure and Setup: The procedure was very similar to the first user study. After the introduction, participants started with a training task. Next, each condition was tested three times and failed authentications were repeated up to three times. *BoD-Gestures* were provided on a piece of paper using graphical versions of the gestures similar to the example in Figure 4.27. The intended side was color coded: red represented *BoD-Gestures*, black represented *Front-Gestures*. After all conditions were tested, participants provided feedback via questionnaire and were compensated with a 5 Euro shopping voucher.

In addition to the two cameras used in the first lab study, we positioned a third camera either to the left or to the right of the participant. The position was derived from the handedness of the user. We made sure that all cameras had perfect sight on the device. The side view was introduced to allow attacks on authentications with *switches* > 0. The camera setup is illustrated in Figure 4.30, right.

Participants: 32 participants were invited using mailing lists, word of mouth and social networks. Nobody had taken part in previous evaluations of *XSide*. The sample was comparable to the first user study. The average age was 25 years (19-38) years. Fourteen participants were female, 18 were male. Most (87.5%) participants owned a touch-based device and the average experience was 3 years (SD = 1.95). 75% of the mobile device users used secure lock mechanisms; eleven used a gesture-based concept.

Usability Findings

The results of the accuracy study indicated that users perform best when both hands are used and interaction takes place in the top area of the screen. Overall, this combination resulted in 102 (9%) errors. In contrast, one-handed interaction at the bottom of the screen resulted in 440 (39%) input errors. As a result, both user studies allowed two-handed interaction. In this Section, we present the efficiency, effectiveness and user perception of both studies.

Efficiency

The results of the first user study are based on 288 authentications per system (1152 overall), the second user study provided 768 samples. As no concept required specific orientation effort, we focus on the input times. Time measurement started with the first touch event and ended with the last touch ("finger up"). The results are based on successful authentications.

Evaluation 2: Concept Comparison As *grid-based gestures* and *BoD-Gestures* did not require explicit confirmation, PIN-times were measured between the touch of the first button and the release of the last button. One sample was removed as authentication time exceeded 25 seconds. The analysis of the video footage confirmed that the participant had been interrupted.

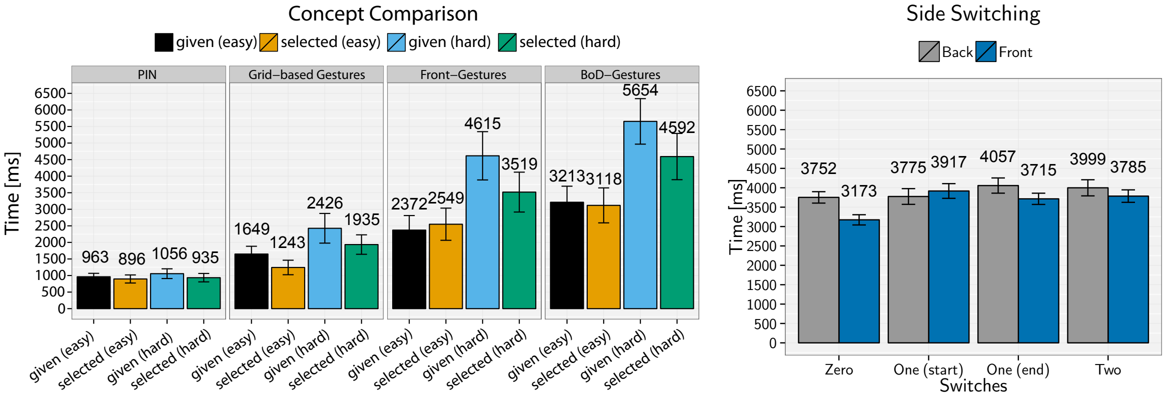


Figure 4.31: The average authentication times of first user study (left) and the second user study (right). The graph on the right indicates that the number of switches has minor impact.

Figure 4.31 (left) illustrates the average authentication times of the tested concepts categorized by *secret-type* and *secret-complexity*. A 4 x 2 x 2 (*system* x *secret-type* x *secret-complexity*) repeated measure ANOVA revealed significant main effects for *system* ($F_{1.93,40.59} = 180.91, p < .001, \varepsilon = 0.64$), *secret-type* ($F_{1.0,21.0} = 16.65, p < .001, \varepsilon = 1.0$) and *secret-complexity* ($F_{1.0,21.0} = 152.79, p < .001, \varepsilon = 1.0$). In addition, we found significant interaction effects for *system x secret-complexity* ($F_{2.21,46.40} = 35.35, p < .001, \varepsilon = 0.74$), *secret-type x secret-complexity* ($F_{1.0,21.0} = 15.00, p < .001, \varepsilon = 1.0$) and *system x secret-type x secret-complexity* ($F_{2.16,45.25} = 5.14, p < .05, \varepsilon = 0.72$).

Bonferroni-corrected post-hoc tests revealed that PIN (Mn = 965ms, SE = 37) was significantly faster than all other systems ($p < .001$). In addition, grid-based gestures (Mn = 1840ms, SE = 93) allowed significantly faster authentication than both *Front-Gestures* (Mn = 3335ms, SE = 173) and *BoD-Gestures* (Mn = 4204ms, SE = 203), all $p < .001$. Finally, front input allowed significantly faster authentication than back-of-device interaction ($p < .001$) and self-selected secrets were significantly more efficient than predefined ones ($p < .05$). This was especially the case, when more complex secrets were used. While the data indicates that *BoD-Gestures* are generally more time-consuming, we found that very efficient authentication is possible. Considering self-selected secrets, the fastest user needed 2.9 seconds on average using a hard gesture and 1.5 seconds using an easy one.

Evaluation 3: Side Switching Figure 4.31 (right) shows the average authentication speed of all tested conditions. A 2 x 4 (*Begin* x *Switches*) repeated measure ANOVA revealed significant main effects for both *Begin* ($F_{1,31} = 14.673, p < .001, \varepsilon = 1.0$) and *Switches* ($F_{2.310,71.601} = 12.18, p < .001, \varepsilon = 0.77$). Furthermore, interaction effects were indicated: *Begin* x *Switches* ($F_{3,93} = 5,438, p < .05, \varepsilon = 0.83$).

The Bonferroni corrected post-hoc tests revealed that authentications were performed significantly faster whenever input started on the front side ($p < .05$). In addition, input without side switches was performed significantly faster than input with side switches ($p < .05$).

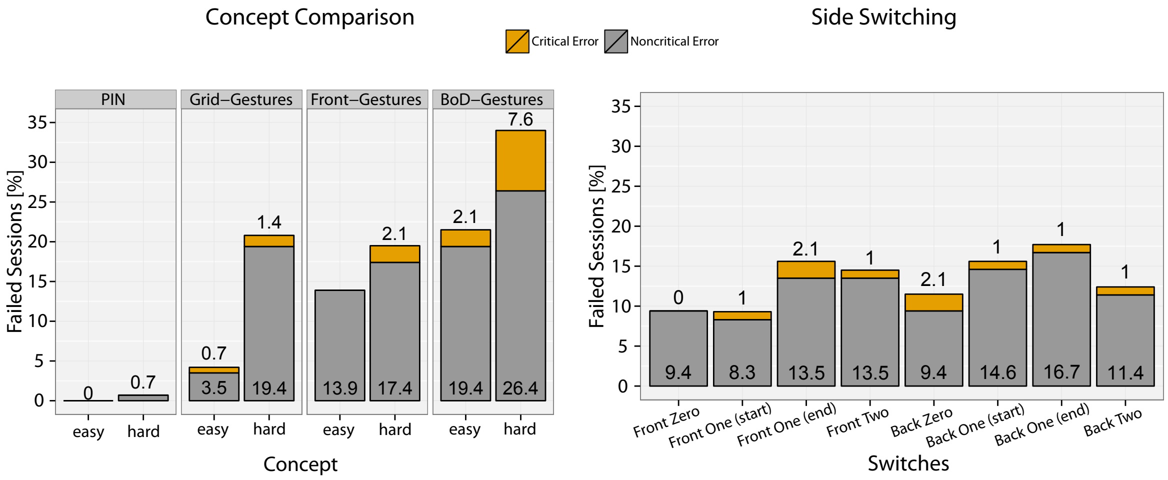


Figure 4.32: The analysis revealed that *Front-Gestures* are as effective as grid-based gestures while performing gestures on the back of the device is more error prone.

Interestingly, while side switches generally add time to the authentication, the contrasts between *one (start)*, *one (end)* and *two switches* were not significant ($p > .05$). As a consequence of the interaction effects, authentication using front only is most efficient.

Effectiveness

Effectiveness is evaluated based on failed authentications. We distinguish between basic errors and critical errors. Basic errors allowed successful authentication after one or two repetitions, while critical errors are based on three failed attempts.

Evaluation 2: Concept Comparison Figure 4.32 (left) illustrates the error rates of all tested concepts categorized by secret complexity. Overall, the error rate is very low for *PIN*. While *Front-Gestures* and *grid-based gestures* perform similar, *BoD-Gestures* are most error-prone.

Even though the number of errors is too small to justify statistical analysis, it indicates that *BoD-Gestures* lead to more errors when complex gestures are used. Overall, 26.4% of such authentication sessions comprised at least one failed attempt and another 7.6% failed completely. To understand the sources of error, we performed a qualitative review based on the taxonomy for the categorization of gesture-based errors which was presented in Section 3.3. All critical errors and 81% of the basic errors could be classified as one of three types:

Additional strokes 29% of the critical and 41% of the basic errors were based on unintentional strokes. In such cases, users accidentally touched the back screen before or after performing the intended gesture.

Mirrored or wrong strokes 50% of critical and 17% of basic errors resulted from mixing up left and right. In such cases, users started the gesture in a wrong direction.

Aborted strokes 21% of critical and 23% of basic errors were based on aborted strokes. In such cases, the input stopped too early as users accidentally lifted the finger or input was performed outside of the touch area.

Evaluation 3: Side Switching Figure 4.32 (right) illustrates the result of the second user study. The overall error rate was 12.1%. Most notably, we observed only nine (1.2%) critical errors. Even though the study was based on a different sample, this indicates that the advanced prototype increased effectiveness. While front-only input was most effective, the data does not indicate a specific impact of side switching.

A qualitative analysis of the failed attempts confirmed the findings of the first study as 39.3% of the input errors were based on mixing up directions, 19.3% resulted from aborted strokes and 14.3% were based on additional strokes. In addition, we found two error types which resulted from the specific study design.

Mixing up sides 20.7 % of the errors were based on mixing up front and back side. Such errors resulted from prescribing the order of side switches. It should be noted that a productive system would support free selection of input sides. Therefore, users mixing up different sides would not be possible in the wild.

False positives 6.4 % of the errors were caused by unexplained touch events. As these events were not visible on the video, we assume that such false positives resulted from the prototype itself: The prototype registered touch events through the 3D-printed cover. We argue that such errors are unlikely to occur with a fully developed device.

Likeability and Perception

In both user studies, participants ranked the system according to usability and satisfaction.

Evaluation 2: Concept Comparison Figure 4.33 (left) illustrates participants' answers to 5-point Likert scales ranging from "fully disagree" to "fully agree". Overall, participants were not in favor of the *XSide* prototype. Only 12% agreed that the concept was efficient, 33% stated it was effective and 29% reported that they liked the system. The usability of the concept was perceived better, when gestures were entered on the front side. In addition, when we asked participants if they would use *XSide* in real life, 13 (54.2%) said "yes". Another two participants acknowledged that they would use the system if simpler gestures were allowed. One participant said "no" as she did not use unlock mechanisms in general. The rest reported that the concept was too cumbersome for practical use.

Evaluation 3: Side Switching Participants' ratings according to side switching are illustrated in Figure 4.33 (right). Overall, users were more positive about the concept. In accordance with the quantitative measures, "Front Only" was rated most effective and most efficient. Furthermore, most participants agreed that using one switch is usable. 25 (78.1%) participants stated they would use *XSide* in the wild. Four users who disagreed would never use any authentication concept on their smartphone.

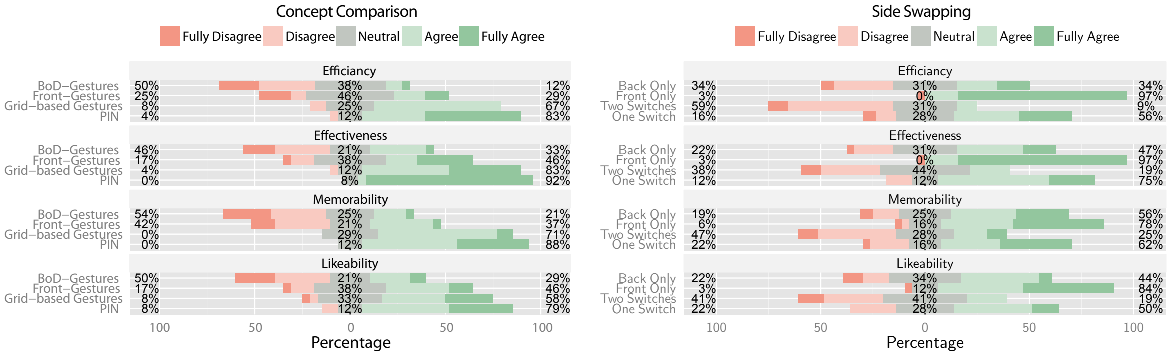


Figure 4.33: The user feedback indicates that users generally prefer front input. However, using *XSide* with one switch provides a good trade-off between usability and security.

Security Findings

Observation attacks were simulated by human attackers who reviewed the video material of the user studies. The attacks were performed assuming a worst case scenario. That is, back-of-device input was attacked from behind, front input was attacked from the front. After watching the authentication, attackers had three guesses. To isolate the effects of observation, we removed any additional cues (e.g., sounds).

Concept Comparison: Attacks were performed by a member of the research team who was familiar with the concepts but had no knowledge of the used secrets. Figure 4.30 gives two examples of the attacker's perspective.

Figure 4.34 (left) summarizes the results. *BoD-Gestures* were most resistant to observation attacks, especially if complex gestures were used. Hard self-selected gestures performed best with only 9 of 24 (38%) identified secrets. In contrast, front-input was very vulnerable independently from the used concept, the secret complexity and the secret type. However, even with PIN and grid-based gestures, some users managed to authenticate in a way that made observation hard. The attacker reported that users entered their secret extremely fast in such cases. Considering *BoD-Gestures*, the attacker mentioned that the viewing angle made it hard to distinguish similar gestures. For example, he often mixed up angled movements like "Left Up" with linear movements like "Left Right" or "Down Up".

Side Switching: In contrast to the first user study the methodology was slightly modified: We increased the number of attackers and provided a monetary incentive. The attacks were performed by four volunteers (two male, two female) who were not part of the user study. We paid a basic salary of 20 Euro and added 30 Cents per successful attack. Each attacker observed eight participants corresponding to one complete cycle of the Latin square design. Before each attack, observers were informed about the order of sides on which interaction will take place (e.g., "front, back, front"). Each attack consisted of three guesses.

Figure 4.34 (right) illustrates the results. With a success rate of 53%, "front only" was most vulnerable, followed by "front (one start)" (38%) and "back only" (31%). Overall,

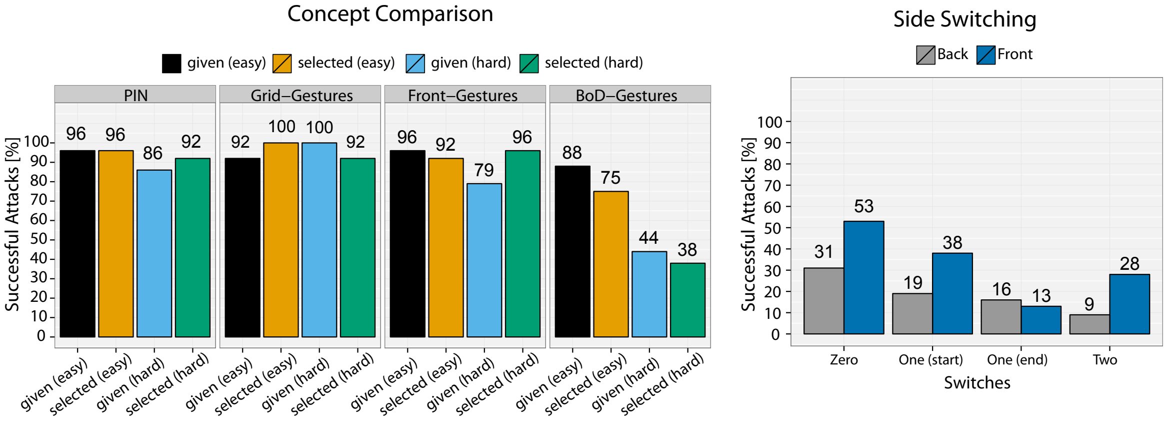


Figure 4.34: The results indicate that *BoD-Gestures* significantly increase observation-resistance. Furthermore, the security level can be adjusted by increasing the number of switches.

the results indicate that security can be improved by starting on the back side of the device. The highest level of security was achieved by starting on the back of the device and then performing two switches (9%). During the attacks, we encouraged attackers to comment on the task. The feedback indicates that slow interaction makes the input easy to observe, while fast inputs were very hard to follow. In addition, observers confirmed the findings of the first study that distinguishing angles on the back side is a difficult task. Besides these general findings, more specific user behavior was mentioned. According to the attackers, some users performed additional finger movements between the gestures. Such movements were hard to distinguish from the actual input and thus increased observation resistance.

Summary

We presented *XSide*, a concept for gesture-based back-of-device authentication. In the first development cycle, we developed the interaction concept, built a basic prototype and compared it to PIN and grid-unlock. The results indicate that *BoD-Gestures* allow for effective and efficient authentication on the back of the device (*H1-1*). The fastest users managed to authenticate within 1.5 to 3.0 seconds. At the same time, the method significantly increased observation resistance (*H1-2*). However, *BoD-Gestures* were less effective and less efficient than the other concepts (*H1-2*). Considering *Front-Gestures*, results were more diverse: While PIN and grid-unlock have been more efficient, *Front-Gestures* performed similarly to grid-unlock in terms of effectiveness. Therefore, we cannot accept hypothesis *H1-4*. Finally, the results revealed that *Front-Gestures* do not provide benefits in terms of security (*H1-5*).

The second development cycle started with the production of an advanced prototype and aimed at evaluating the impact of side switching. We hypothesized that side-switching might empower user to react more flexible to observation risks. The results indicated that side switching increases observation resistance. However, we reject hypothesis *H2-2* as one side switch resulted in similar security improvements as adding two side switches. Moreover, even though front-only input was most efficient, the results did not indicate that higher num-

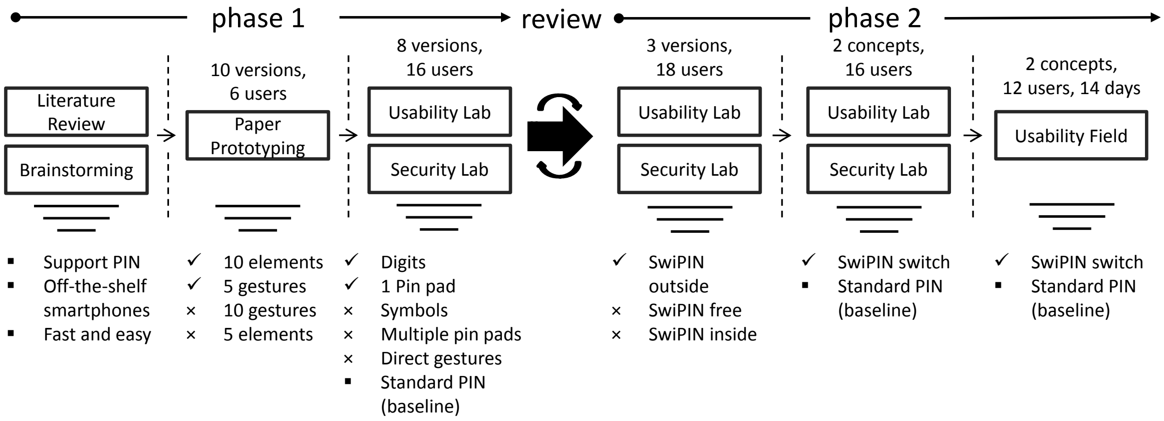


Figure 4.35: Concept development was based on two phases: In the first development cycle, we developed the general concept and compared it to PIN. In the second phase, *SwiPIN* was improved, implemented as fully functional prototype and finally tested in the field.

bers of switches generally reduce efficiency and effectiveness (*H2-2*). Overall, improving the prototype already resulted in faster authentication, reduced error-rates and increased user acceptance. Therefore, we are confident that the performance and the acceptance of *XSide* will further improve as soon as it is more naturally integrated in real customer devices.

4.3.5 SwiPIN: Utilizing Gestures to Protect PIN-Entry

SwiPIN was developed as low-fidelity and high-fidelity prototypes and tested in the lab and in the field. Figure 4.35 illustrates important milestones of the design and evaluation process. The process resulted in a gesture-based authentication mechanism which allows efficient, effective and observation-resistant authentication on off-the-shelf mobile devices. While the first design phase was indispensable to elaborate the concept and evaluate different design alternatives, we will mainly focus on the second phase which revealed interesting aspects concerning user behavior and the feasibility of the concept itself. We will first present the evaluation strategy and outline the designs of the different user studies. The second part of this Section presents the results of the performed user studies in chronological order.

Prototyping and Evaluation Strategy

The *SwiPIN* concept resulted from testing various design alternatives. The main aspects of the preliminary development process will be summarized in the first part of this Section. The second part focuses on the evaluation strategy of the main user studies.

Evaluation 1: Paper Prototyping and Preliminary User Study

The development started with an initial brainstorming which aimed at finding efficient and effective gesture-based interaction concepts that allow observation-resistant authentication

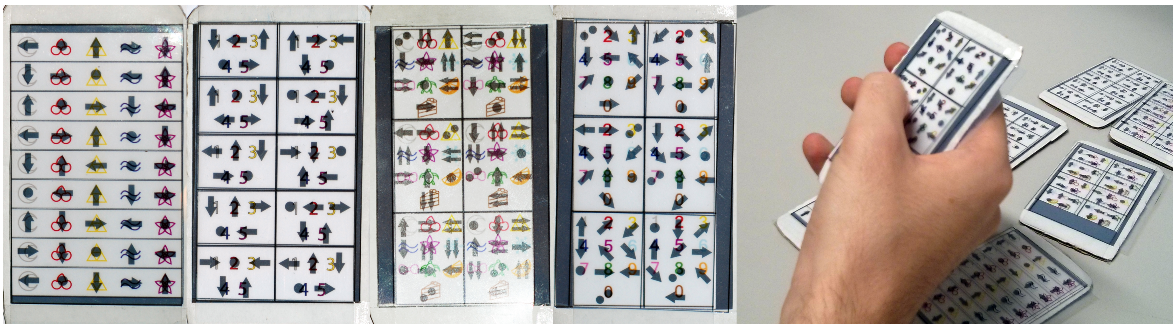


Figure 4.36: Different design options were tested in a first user study using paper prototypes. The study focused on the characteristics of the gesture-set and the spatial layout of the input elements.

on off-the-shelf mobile devices. The ideas included to transfer existing methods for desktop computers (e.g., [278]), use the built-in camera to establish a secret channel and various other methods which are based on indirect input or distraction. After a review of all concept ideas, using randomly assigned swiping gestures was identified as the most promising concept. The concept was further developed in a prototyping study and a preliminary lab study.

Paper Prototyping: The idea was concretized in a preliminary lab study using low-fidelity prototypes. Based on ten different paper prototypes, we informed the following design decisions in terms of gesture sets, mappings and layout options:

Gesture Set We tested three different gesture sets: 1) the basic set including five gestures (up, right ,down, left, tap), b) the multi-touch set including ten gestures (basic set + basic set using two fingers), 3) the diagonal set including ten gestures (basic set + right-up, right-down, left-up, left-down, double tap).

Number of Input Elements We tested two different sets. One set comprised 1) five input elements, the other set comprised (2) ten input elements.

Secret Type We tested two types of secrets: (1) digits and (2) concrete symbols.

Number of Redundant Fields We tested different layouts comprising (1) six, (2) eight, and (3) ten input fields. An input field describes a complete set of input elements.

Figure 4.36 illustrates examples of the tested paper prototypes which were tested by six participants. We collected qualitative data via semi-structured interviews and a questionnaire.

Preliminary Lab Study: As the paper prototyping indicated that users desired both the basic gesture set and ten input elements we opted to subdivide each input field in two sectors. The design decision was kept until the final prototype. In contrast, no clear preferences for the number of fields were found and it remained unclear if users prefer digits or symbols. To shed light on these important questions and to gather preliminary insights into the usability and into the security of the concept, we performed a high-fidelity user study. For this purpose, the concept was implemented for Android and evaluated in the lab. The user study followed a repeated-measures design. The independent variables were *field number* with four levels (*zero*, *one*, *four*, *six*) and *secret type* with two levels (*digit*, *symbol*). *Field number = zero* represented the two baseline conditions: *PIN* and *SymbolTap*, a version of PIN using symbols. The order of the eight conditions was counterbalanced using a Latin square design.

We recruited 16 experienced touchscreen users, ten were male, six female. The participants' average age was 29 (23-59). Each secret was entered five times, three attempts were granted per authentication. Quantitative performance data was collected via built-in logging mechanisms. Qualitative feedback was collected with questionnaires. In addition, the interaction was filmed. The video footage was used for the later security analysis.

Evaluation 2: Starting Point Lab Study
The preliminary evaluation led to the concept of *SwiPIN* but revealed various areas for improvement. Most critically, user behavior had jeopardized observation resistance as users often started the gestures on the intended buttons. As a consequence, we revised the concept in the second development phase and modified *SwiPIN* in a way which enforced desired user behavior (indirect gesture input). In addition, we rejected the idea of using multiple redundant input fields and decided to discard the use of symbols to support seamless integration into common PIN concepts. Figure 4.28 illustrates three modifications of the redesigned *SwiPIN* concept. *SwiPIN* "free" (left) resulted from the first development phase and served as baseline, *SwiPIN* "inside" and *SwiPIN* "outside" were designed to enforce indirect gestures. The three versions of *SwiPIN* were evaluated in a repeated-measures lab study.

Hypotheses: Two (main) hypotheses were defined for the first lab study.

H1-1 All layouts of *SwiPIN* are equal in terms of effectiveness, efficiency and likeability.
H1-2 Indirect gesture input does significantly increase observation resistance.

Design: The study followed a repeated measure within participants design. The only independent variable was *layout* with three levels (*free*, *inside*, *outside*). The order of *layout* was counterbalanced. Entered PINs were randomly generated and comprised four unique digits.

Procedure and Setup: Each session started with an introduction to shoulder surfing threats. We told participants to enter the digits as fast and as error-free as possible. Each concept was evaluated using the following procedure: a) Explanation of the functionality and training. b) Authentication with three different PINs. Each PIN was entered five times, failed attempts were repeated up to three times. c) Users rated the concept. After all systems were tested, participants answered a final questionnaire comparing all systems.

Expected PINs were communicated with alert dialogs. As soon as the dialog box was dismissed, authentication started and quantitative data was collected using built-in logging mechanisms. In addition, the interaction was filmed for the later security evaluation. Finally, participants were compensated with a 5 Euro shopping voucher.

***Participants**:* We recruited 18 participants via the university mailing list and social networks. The average age was 25 years (20-32), thirteen were male, eight female. All participants reported to use touch-based smartphones on a daily basis.

Evaluation 3: Feasibility Lab Study

The starting point lab study indicated that *SwiPIN* "outside" performed best in terms of usability and security. Therefore, this layout was specified as the final *SwiPIN* concept and evaluated in a feasibility study. The study had two main goals: 1) It aimed at comparing the final SwiPIN concept to traditional PIN-entry, 2) We investigated the feasibility of switching between PIN and SwiPIN.

***Hypotheses**:* Four (main) hypotheses were defined for the feasibility study.

H2-1 *PIN* is more efficient and more effective than *SwiPIN*.
H2-2 *SwiPIN* is more resistant to observations than *PIN*.
H2-3 *SwiPIN* is perceived to be sufficiently efficient and effective.
H2-4 *SwiPIN* allows seamless integration into *PIN*.

***Design**:* The study was based on a repeated-measure within participants design. The independent variable was *system* with two levels (*SwiPIN*, *PIN*). SwiPIN was based on the final layout of SwiPIN "outside", PIN was inspired by implementation of current smartphones. *System* was counterbalanced.

***Procedure and Setup**:* While the procedure followed the main steps of the previous study, there were three modifications: 1) Only one PIN was tested per condition, 2) We added a *PIN* training task and 3) We added a concept switching task. After the introduction, *SwiPIN* and *PIN* were tested separately using the following procedure: a) Concept training followed by PIN training. While the concept training allowed the participants to get familiar with the prototype using arbitrary digits, the PIN training helped the user to adapt to the assigned secret. For this purpose, the secret was entered ten times in a row. Next, b) the same PIN was used to authenticate five times. After both concepts had been used separately, we performed a switching task. For this purpose, a new *PIN* was assigned and trained ten times. Afterwards, users authenticated five times using *PIN*, five times using *SwiPIN* and again five times using *PIN*. Switching between *PIN* and *SwiPIN* was accomplished using a software button which was positioned as the top-right corner of the screen.

Qualitative data was again collected via questionnaire, quantitative data was gathered using logging mechanisms. The interaction was filmed, participants were compensated with a 5 Euro shopping voucher.

***Participants**:* We recruited 16 participants via a university-wide mailing list. Participants' average age was 27 years (21-37), twelve were male, four female. All participants reported to be experienced smartphone users. In addition, eight persons had taken part in the first user study. However, no significant differences were found between novel users and users with prior *SwiPIN* experiences.

Evaluation 4: Longitudinal Field Study

The next Section presents the results in detail. In summary, the feasibility study had indicated that *SwiPIN* allows efficient, effective and observation resistant PIN-entry. However, we aimed at testing the feasibility of the concept in the wild. For this purpose, we performed a final field study.

Hypotheses: Four (main) hypotheses were defined for the field study.

H3-1 *PIN* is used for most authentications.
H3-2 *SwiPIN* is used whenever observation risks are perceived.
H3-3 *SwiPIN* is perceived as sufficiently efficient and effective for daily authentication.
H3-4 *SwiPIN* is a useful supplement to *PIN*.

Design: The evaluation was based on a longitudinal field study. Users installed *SwiPIN* on their personal devices and used it as primary authentication mechanism for 14 days. We applied an experience sampling approach similar to the procedure reported in Section 3.2. That is, we presented short questionnaires to inform the user's context. As we aimed at observing natural user behavior, the switching between *SwiPIN* and *PIN* was not controlled.

Procedure and Setup: The prototype of the feasibility study was implemented as lock screen replacement. We made sure that the application worked with different screen sizes and manufactures. To ensure a familiar user experience, we integrated a clock and replicated the common lock screen behavior. For example, users were able to answer calls without authentication. Concept switching was again triggered with a software button which was placed at the top-right corner of the screen. The concept allowed three failed attempts before a fallback PIN was required.

Participants were selected based on their answers to an initial questionnaire. The survey collected basic demographic information and ensured that the user's smartphone met the requirements of the study (e.g., Android v. 4.0+). On the first study day, participants were invited to the lab. We helped with installing the software, ensured its functionality and explained the procedure of the study. Participants then used the concept over the course of two weeks.

We collected quantitative data using built-in logging mechanisms. We logged common performance data as well as the number of digits used for the PIN. This information was required to interpret authentication times. However, we did not record the PIN itself. As already mentioned, qualitative data was collected using an experience sampling approach. For this purpose, we presented a short questionnaire 30 minutes after *SwiPIN* had been used. The questionnaire comprised four questions concerning 1) the current situation, 2) the reason for using *SwiPIN*, 3) the perceived usability and 4) the perceived security. While the first

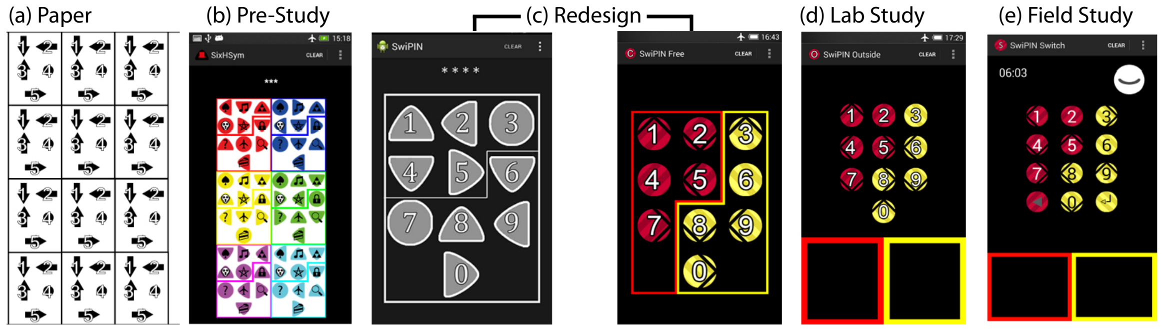


Figure 4.37: *SwiPIN* was gradually improved over the course of the project.

two questions presented common options (e.g., "at home" or "I felt observed") and allowed text input, the last two questions were based on the relative positioning of a slider. None of the questions presented default answers. Users were allowed to postpone the questionnaire by clicking a software button.

After the two weeks, participants returned to our office for a debriefing. We uninstalled the application, collected the log files and asked participants to fill in a final questionnaire. Finally, participants were compensated with a 20 Euro shopping voucher.

Participants: We recruited twelve participants via mailing list and word-of-mouth advertising. Unfortunately, four participants had to be excluded due to hardware problems. Therefore, the final sample was based on eight participants, among them three females and five males. The average age was 28 years (19 - 33).

Usability Findings

The previous Section presented the used prototypes and provided details on the design of the performed user studies. This Section presents the results of the evaluations in chronological order. We illustrate the evolution of *SwiPIN* from its first interactive version to a real lock screen replacement. Each developmental stage provides valuable insights which show the importance of a systematic design approach based on multiple iterations.

Efficiency

Efficiency was assessed using the mean authentication time. As recommended in Section 4.2, time measurement was again split into orientation phase and input phase. Figure 4.38 illustrates the average authentication times observed in the performed lab studies. As the efficiency data of the field study can only serve as rough approximation, it was excluded from this analysis. The results are based on correctly entered PINs, the data was normally distributed and allowed for parametric tests.

Evaluation 1: Preliminary User Study The preliminary user study aimed at analyzing the effects of redundant input fields using symbols and digits. As mentioned in the previous

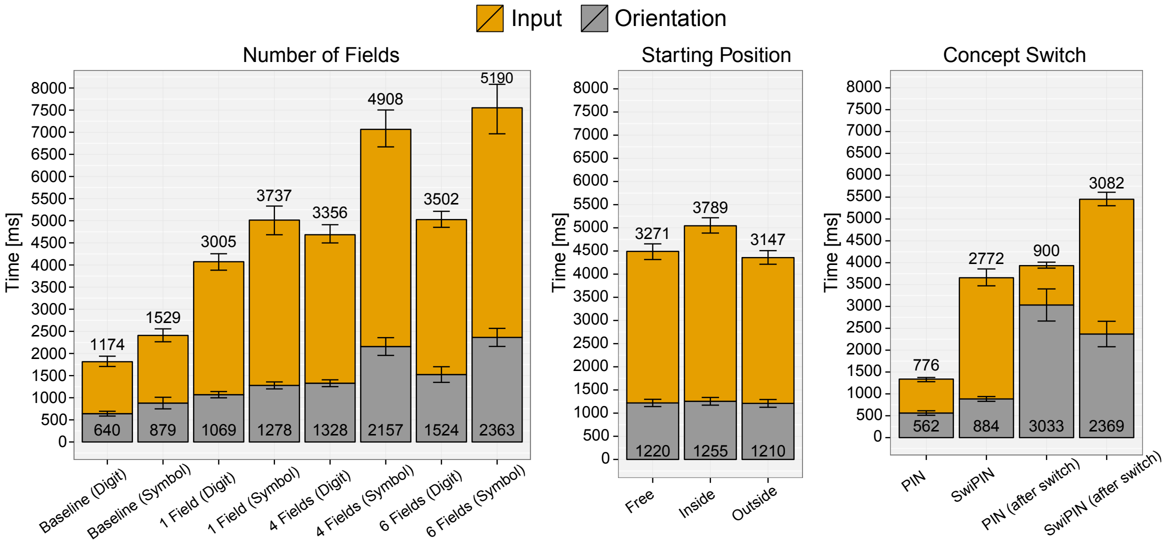


Figure 4.38: Efficiency of *SwiPIN* was gradually improved over the course of the project.

Section, we tested three versions with one, two and six input fields and compared them to PIN and symbol-based PIN. The results are based on the average authentication times of the fourth and the fifth run as initial analysis indicated unbalanced performance in the first three runs. The results are depicted in Figure 4.38 (left).

A repeated measure ANOVA revealed a significant impact of *field number* on orientation time, $F_{3,45} = 30.62, p < .001$. Bonferroni corrected post-hoc tests indicate that both baseline approaches allowed significantly faster orientation times, $p < .05$. In addition, concepts with one input field supported significantly faster orientations than concepts with four and six fields, $p < .05$. Furthermore, *secret type* significantly influenced orientation times, $F_{1.00,15.00} = 34.35p < .001, \varepsilon = 1.0$. The known layout using digits performed significantly faster than the layout based on symbols, $p < .001$. The effect became even stronger when more fields were present.

A repeated measure ANOVA comparing the average input times showed similar results. Again, the *number of fields* had a significant effect on the input times, $F_{3,45} = 59.09, p < .001$. Again, both baseline concepts performed significantly faster than all other layouts ($p < .001$) and input using one field was significantly faster than input based on four or six fields, $p < .05$. In addition, *secret type* significantly influenced input speed, $F_{1.00,15.00} = 21.82p < .001, \varepsilon = 1.0$. Input was significantly faster when digits were used, $p < .001$. The effect was stronger, when multiple input fields were present.

In summary, the preliminary user study indicated that *SwiPIN* was most efficient when a known layout (i.e., PIN layout) was used and when interaction was based on one input field. In this case, users needed about 4.1 seconds to authenticate.

Evaluation 2: Starting Point Lab Study After the redesign of *SwiPIN*, we ran the second user study to investigate the impact of different starting positions. We again excluded the first three runs of each *PIN* × *layout* combination. As each participant contributed three PIN entries, the results are based on the average of three authentications measured in the last two runs. The results are illustrated in Figure 4.38 (center).

A repeated measure ANOVA found no significant main effect of starting positions on orientation times, $p > 0.05$. However, a repeated measure ANOVA analyzing the average input times found a significant main effect for *layout*, $F_{1.44,24.50} = 13.48, p < 0.001, \varepsilon = 0.72$. Bonferroni corrected post-hoc tests showed that forcing users to start *inside* of the PIN pad resulted in significantly slower input times, $p < 0.05$. In contrast *SwiPIN free* and *SwiPIN outside* allowed similar input speed.

In summary, the results indicated that forcing users to start their gesture outside of the PIN pad was most efficient and allowed users to authenticate within 4.3 seconds on average.

Evaluation 3: Feasibility Lab Study Using SwiPIN *outside*, the feasibility study aimed at analyzing the effects of switching between *PIN* and *SwiPIN*. We assessed the efficiency of *PIN* and *SwiPIN* using the last two runs. In addition, the effects of concept switching were evaluated using the last input before the switching event and the first input after the switch. Figure 4.38 (right) illustrates the results including the observed authentication times after performing the concept switch.

We compared the average orientation effort and the average input times of *PIN* and *SwiPIN* using two-tailed dependent t-tests. The results indicate that *PIN* demands significantly shorter orientation phases ($t_{15} = -4.29, p < 0.05, r = .74$) and at the same time supports significantly faster input times, $t_{15} = -10.38, p < 0.001, r = .94$. Therefore, using traditional *PIN* saves 2.3 seconds on average. A repeated measure ANOVA comparing the orientation times of *PIN* and *SwiPIN* before and after switching the concept revealed that both times are significantly increased after switching the concept, $F_{1.00,15.00} = 70.49, p < 0.001, \varepsilon = 1.0$. However, input times are not affected by the event of concept switching, $p > 0.05$.

In summary, the results of the feasibility study revealed that *PIN* enables significantly faster authentication than *SwiPIN*. At the same time, seamless integration of *SwiPIN* and *PIN* seems feasible even though users required slightly more time.

Evaluation 4: Longitudinal Field Study The field study aimed at understanding user behavior and at assessing the real-world feasibility of *SwiPIN*. As the feasibility study already indicated that efficiency drops after concept switches and as the data set is small, the explanatory power of the measured authentication times is limited. The results are based on the performance of seven *SwiPIN* users. Analogue to the lab studies, we used only successful authentications and excluded outliers.[10]

[10] We utilized the average authentication time observed in the feasibility study to identify outliers. The observed lab study times were multiplied by three to set the cutoff for the field study. For *PIN*, we excluded 5.9%

The participant, who used *SwiPIN* most often was also most efficient. The person needed 1165 ms per digit which corresponds to 4.5 seconds for a four-digit PIN. The overall average is based on 4439 *PIN* authentications and 53 *SwiPIN* authentications. Using *PIN*, participants needed an average of 415 ms per digit (SE:2, 180-1000), using *SwiPIN* 1563 ms per digit (SE:68, 812-2672) were used. Compared to the values of the feasibility study, this indicates that interaction in the wild reduced the authentication speed by 24% (*PIN*) and 58% (*SwiPIN*), respectively.

In summary, the field study indicated that *PIN* and *SwiPIN* perform slower when used in the wild. At the same time, the results were promising as participants who used *SwiPIN* frequently were able to authenticate efficiently (i.e., 4.5 seconds).

Effectiveness

Effectiveness was assessed by analyzing the number of input errors and corrected attempts. The results are based on the same data as used for the efficiency analysis.

Evaluation 1: Preliminary User Study Both baseline approaches were easy to use: PIN was most effective as we observed no input errors. When using symbols in the baseline, one out of 32 (3.1%) authentications failed. Using *SwiPIN* with one field was still effective, as we observed only two errors (6.3%), independently from the used *secret type*. However, increasing the number of input fields did also increase the number of input errors. Using four fields, 18.8% of the authentications failed in the numeric condition and 12.5% failed using symbols. Using six input fields, 12.5% of the authentications failed in both conditions. In summary, the results indicated that *SwiPIN* is effective, when only one input field is used. Increasing the number of input fields made authentications more error-prone.

Evaluation 2: Starting Point Lab Study Overall, the error rates were low. Five out of 108 (4.6%) authentications failed using *SwiPIN free*, six (5.6%) attempts failed using *SwiPIN outside*. *SwiPIN inside* was most error-prone with twelve errors (11.1%). In summary, the results indicated that forcing users to start their gestures *outside* of the PIN pad did not have any negative impact on effectiveness.

Evaluation 3: Feasibility Lab Study When testing *PIN* and *SwiPIN* separately, the error rates were low. We observed one failed attempt (3.1%) in each condition. Overall, switching between the two concepts was effective as well. Nevertheless, while switching from *SwiPIN* to *PIN* was error-free, two participants failed within their first attempts on *SwiPIN*, after *PIN* had been used.

Evaluation 4: Longitudinal Field Study Compared to the lab environment, the field study revealed higher error rates for *SwiPIN*. While 143 of 4582 (3.1%) sessions failed when *PIN* was used, 8 of 61 (13.1%) attempts failed using *SwiPIN*. The results indicate that *PIN* is

of the samples, for *SwiPIN*, 34.6% were cleaned. The participants' feedback confirmed that such long authentication sessions resulted from interruptions or from presenting the concept to others.

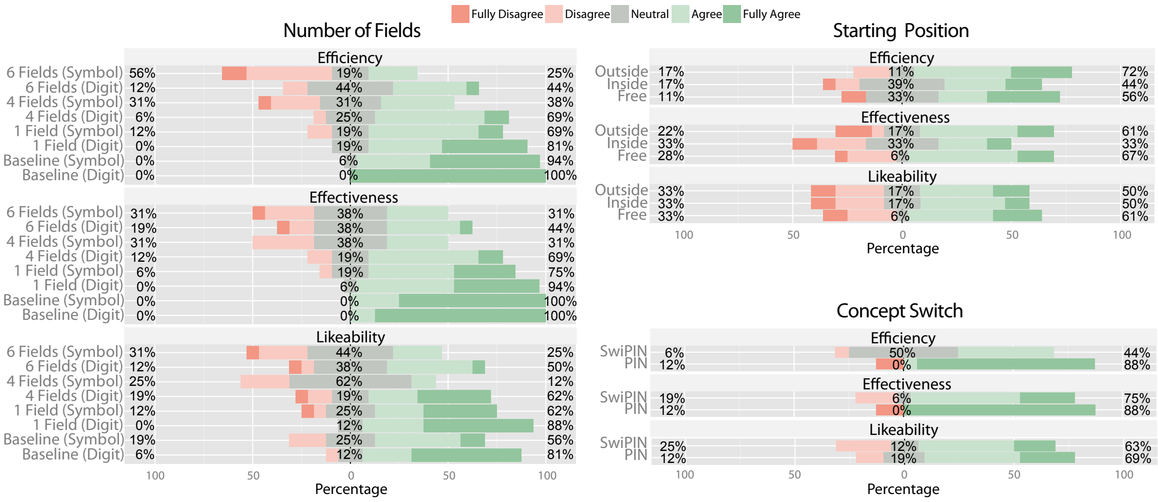


Figure 4.39: Perception ratings of the different lab studies. Participants preferred the use of one input field. Overall, participants were most positive about *SwiPIN* outside.

easier to use than *SwiPIN*. However, the data set of *SwiPIN* is very small and has limited explanatory power.

Likeability, Perception and User Behavior

While the lab studies provided preliminary insights into user acceptance and perception, the field study helped to understand user behavior and the context of use. Figure 4.39 illustrates the user ratings of the lab studies which were collected using 5-point Likert scales.

Evaluation 1: Preliminary User Study The baseline approach was rated best in terms of efficiency and effectiveness. However, *SwiPIN* was liked best, when one field and digits were used. In this regard, none of the participants disagreed that *SwiPIN* with one field and digits was effective and efficient. Overall, participants were not in favor of using multiple fields or symbols. We furthermore asked participants if they would use *SwiPIN* on their personal device. While no participant was willing to use the six-field version, the one field version was accepted by 14 out of 16 participants.

Evaluation 2: Starting Point Lab Study Most participants agreed that *SwiPIN free* and *SwiPIN outside* were efficient and effective. In contrast, *SwiPIN inside* was rated slightly lower for both aspects. While *SwiPIN outside* was rated slightly more efficient than the *free* version, more users reported that they would prefer free starting points. However, in a direct comparison, most users (61%) preferred *SwiPIN outside*. While *Free* was favored by 39% and thus scored second best, *SwiPIN inside* was not selected. Finally, 14 out of 18 (78%) reported that they would use *SwiPIN* in the wild. Overall, forcing users to start their gestures *outside* of the PIN pad did not hamper user acceptance.

Evaluation 3: Feasibility Lab Study While *PIN* was rated very good in terms of efficiency and effectiveness, *SwiPIN* was rated acceptable in terms of efficiency and good in terms of effectiveness. However, most participants stated to like both *PIN* and *SwiPIN*. When asked if they would use *SwiPIN* in daily life, 12 out of 16 (75%) participants acknowledged they would use it as primary authentication system. Furthermore, all participants indicated that they would use *SwiPIN* as an add-on to their primary lock screen. We additionally asked if using *SwiPIN* could communicate mistrust. However, only one participant agreed and most participants strongly disagreed.

Evaluation 4: Longitudinal Field Study In contrast to the self-reported data of the lab evaluations, the field study provided insights into the actual user behavior. On average, 2% (range: 0-5%) of the logged authentications were performed using *SwiPIN*. In absolute values this means that 95 of 4997 authentications were performed using the secure interaction method. On subject level, *SwiPIN* was used between zero and 29 times and *PIN* was used between 29 and 1270 times. This indicates that the sample included various user types ranging from infrequent users to power users. As a consequence, participants contributed a different number of mini questionnaires. Therefore, the data was summarized on subject level before it was analyzed. Using the 10-point scales, participants indicated a median of 7 (range: 5-10) concerning usability and a median of 9 (range: 5-9) concerning security.

Furthermore, participants reported that most of the *SwiPIN* input was performed at home (54%), followed by public transport (18%), other (18%) and work or school (10%). We logged only 13 instances, where participants actually felt observed, nine (69%) of them in public transport. The rest of the *SwiPIN* input was performed due to other reasons: 53% of the instances were tagged with “for no reason”, 19.0% were performed to show *SwiPIN* to someone else and 11% had “other” reasons. This indicates that the results are biased by novelty effects. The results of the final interview indicate that the concept was indeed well accepted but user experience was downgraded by implementation flaws. Several participants complained that the system lagged. In addition, some participants had to confirm two lock screens (*SwiPIN* and the native mechanism) to authenticate. The feasibility of the concept was finally indicated by the anecdotal fact that one participant continued using *SwiPIN* on her personal device. Six weeks after the study was finished, we discovered that the application was still sending log data.

Overall, the results of the field study indicate that *SwiPIN* would be accepted if it was natively implemented. Furthermore, the observations confirmed the findings of Section 3.2 that shoulder surfing risks are rarely perceived and most of the authentications are performed in trusted environments.

Security Findings

The security analysis was based on the simulation of cognitive observation attacks. For this purpose, the video records of the user studies were cut into single successful attempts and reviewed by experienced *SwiPIN* users. Each attack consisted of three attempts while each condition was attacked once using the first successful authentication event.

Evaluation 1: Preliminary User Study The study aimed at analyzing the security benefits of multiple input fields. Overall, 81% (n = 78) of secrets were successfully observed when direct taps were used (baseline condition). Using *SwiPIN* with one input field allowed to successfully observe 40% (n = 38) of the secrets. When four fields were present 15% (n = 14) of the secrets were exposed and with six fields 7% (n = 7) of the authentications were successfully attacked. While this indicates that providing multiple input fields improves security, we found that this was only the case when users actually used multiple fields for their input. However, the analysis reveals that 35 out of 64 (55%) password inputs were based on only one field even though multiple fields were present. 95% (n = 20) of the successfully attacked authentications were based on this behavior. In contrast, only one authentication was successfully attacked when multiple fields were used for authentication. Concerning the secret type, the results indicate that more secrets were stolen using symbols as the unfamiliar layout led to slow interaction which was easier to observe.

In addition to the quantitative analysis, we performed a qualitative analysis of the video footage to understand why observation attacks were successful. The analysis revealed that *SwiPIN* was often rendered ineffective by obvious user interaction. Many users revealed the entered token by starting their gesture on the respective input element. Please note that gestures can actually start anywhere in the respective half of the display. Another common behavior was to hover over the intended button before performing the gesture. Even though we found that increasing the number of input fields reduces the effect due to the smaller areas, the problem was still present.

In summary, the user study confirmed that *SwiPIN* can effectively prevent observation attacks but specific user behavior makes it vulnerable to observation attacks.

Evaluation 2: Starting Point Lab Study The study aimed at analyzing the security advances of predefined starting points. This time, the security analysis was performed by three attackers (1 female). Each of the attackers observed 54 authentications. The results are based on binary success (true/false) and the relative success rate. The relative success rate is described as the overlap of the guessed secret and the entered secret.

SwiPIN outside was most secure as only one PIN was successfully attacked (binary: 1 of 54; overlap: 35.6%). In contrast, enforcing user to start their gestures *inside* of the PIN-pad led to five successful attacks (binary: 5 of 54; overlap: 44.4%). A qualitative review revealed that slow interaction enabled attackers to observe parts of the mapping and reduce the search space for their guesses. As indicated by the preliminary security study, allowing *free* starting points makes the system prone to observation attacks (binary: 8 of 54; overlap: 49.5%). The video review confirmed that participants again started their gestures on the intended digit. Therefore, the results confirmed that *SwiPIN outside* performed best in terms of security.

Evaluation 3: Feasibility Lab Study Finally, we again compared *SwiPIN* to traditional *PIN*. The baseline results confirmed the findings of the preliminary user study as PINs were correctly identified in almost all cases (binary: 14 of 16; overlap: 92.2%). Nevertheless, two participants managed to enter their PINs fast enough to prevent successful observation. In contrast to *PIN*, *SwiPIN* was significantly more resistant to observation attacks (binary: 2 of 16; overlap: 35.9%). The two successful attacks were based on the fact that attackers managed to observe parts of the input and guessed the rest. For example, one attacker noted that he had observed the sequence of colors: red, yellow, red and yellow. This gives a chance of $(\frac{1}{5})^4 = 0.16\%$ to guess the correct PIN. In summary, the results confirmed that *SwiPIN* is significantly more secure against observation attacks than *PIN*.

Summary

We presented *SwiPIN*, a concept which allows secure PIN-entry based on simple touch gestures. We reported the systematic design approach and illustrated how *SwiPIN* was successively improved from the first idea to a working prototype. The Section illustrated the importance of early user testing as the security of *SwiPIN* was significantly influenced by user behavior. We learned that providing multiple input fields could theoretically increase observation resistance but that such features are unlikely to be used in the wild. In addition, the results showed how specific user behavior can directly downgrade security. While preliminary designs allowed free starting points for gestures, the improved version enforced more secure behavior.

The evaluation indicated that *SwiPIN outside* was as usable as the *free* version (*H1-1*) but significantly more secure (*H1-2*). At the same time, the relation of input elements and output elements was important as enforcing gestures on the insight of the PIN pad turned out to be a bad idea. Indeed, the final layout of *SwiPIN* was well accepted (*H2-3*), highly secure (*H2-2*) and showed good performance. After some minutes of training, users already achieved an average authentication speed of 3.7 seconds and low error-rates (3%). Nevertheless, traditional PIN-entry still performs better in terms of usability (*H2-1*). As a consequence, we suggest to provide *SwiPIN* as a secure add-on to more efficient mechanisms. The last lab study proved the feasibility of an integration of *PIN* and *SwiPIN* (*H2-4*). In addition, when *SwiPIN* was tested in the field, participants were very positive about the concept (*H3-3*) and agreed that it is a useful supplement to *PIN* (*H3-4*). Finally, *SwiPIN* was used as intended as participants used traditional PIN-entry for most authentications (*H3-1*) and switched to *SwiPIN* when required (*H3-2*).

Besides presenting a usable and secure authentication mechanism which is suitable for daily use, the Section contributed by providing general insights: Firstly, we illustrated how a systematic design process helps to develop a feasible solution. Secondly, we showed how using gestures can improve traditional authentication mechanisms and how this allows seamless integration of efficient and secure interaction models.

4.3.6 Discussion and Implications

In this Section, we summarize the main findings concerning *XSide* and *SwiPIN* and discuss implications for observation-resistant gesture-based authentication concepts.

Gestures Allow Observation-resistant Authentication but Output Matters

The results show that touch gestures are well suited to build usable and secure authentication mechanisms. Both *XSide* and *SwiPIN* allow reasonably fast authentications which are well below the reported times of most related work (see Section 2.3.3). In addition, both concepts are significantly more resistant to observation attacks than currently used methods like PIN and grid unlock. However, it is important to note that the gesture itself is not observation resistant but security is achieved by a clever combination of various design factors. In this regard, the configuration of output elements represents the most important design factor. While both systems provide aggregated *feedback*, the utilization of *guidance* is completely different and illustrates two potential approaches.

XSide achieves observation resistance by forgoing visual guidance. Since target-oriented gestures are not usable in this context, self-contained gestures were provided. As a consequence, the system naturally supports eyes-free interaction. Eyes-free interaction allows more secure authentication but may also provide usability benefits as users do not have to focus on the device. In addition, *XSide* provides observation resistance without randomization and thus allows motor learning. In contrast to *XSide*, the security concept of *SwiPIN* is specifically based on visual guidance. While the spatial arrangement of input elements is fixed, output elements are randomly positioned and need to be recognized during authentication. Indeed, the concept appears random to observers but quasi-static to users. The results indicate that participants used their spatial memory to keep search times short as they performed significantly faster when a known layout (i.e., PIN) was used.

We conclude that *output elements* and *interaction style* are the most important factors when dealing with observability. In addition, we suggest to minimize randomization. For example, both systems benefit from fixed spatial positions of *input elements*.

Traditional Concepts can Benefit from Gesture-based Interaction

As already indicated, the design of *SwiPIN* and *XSide* also illustrates how the same gestures (up, down, left, right) can be utilized in different ways. While *XSide* implements a new secret type, *SwiPIN* utilizes such gestures to protect PIN-entry.

As popular grid-based interaction concepts are not feasible for back of device authentication, it was an appropriate decision to design a novel type of secrets which would not require accurate pointing and dragging. However, introducing new password concepts can have various disadvantages. Unfamiliar concepts may alienate users and hinder adoption. Furthermore, new secret types may require modified infrastructures (e.g., data bases). *XSide* was specifically designed to minimize such negative effects. Firstly, the provided shapes were similar to widely accepted drawmetric concepts (e.g., Android unlock). Secondly, we made sure that no further modifications were required: *XSide* gestures can be translated and stored as alphanumeric strings. In contrast, *SwiPIN* shows how changing the *interaction style* can improve the security of traditional concepts and illustrates that minimal modifications of the user interface have a significant impact. We argue that minimizing the differences to established concepts can increase user acceptance. In addition, such concepts can be utilized in other domains (e.g., ATMs) without the need of cost-intensive changes in the infrastructure.

We conclude that novel systems benefit from supporting concepts and secrets which are already familiar to the user. However, introducing novel secret types is justified if significant benefits can be expected which cannot be provided with traditional solutions (i.e., eyes-free interaction).

Adaptive Interaction Concepts can Increase Usability and User Acceptance

Observation-resistant concepts often introduce more complex interaction tasks. For example, switching sides naturally adds a certain amount of time. Similarly, the use of indirect gestures takes more time than directly tapping the respective buttons. At the same time, Section 3.2 revealed that most authentications are performed in trusted environments where increased security is unnecessary. This aspect was taken into account by designing adaptive interaction concepts which allow to increase observation resistance on-demand.

Both *XSide* and *SwiPIN* allow users to adapt the security level to the current environment. The evaluation of *XSide* showed that switching sides increases observation resistance. However, in some situations, it might be beneficial to use only one side of the device for input. Specifically, as long as no threats are present, users can authenticate using only the front side. Front-only input was perceived efficient and was shown to be faster and less error-prone than other conditions. Similarly, *SwiPIN* was designed to be used as an addition to traditional PIN-entry. As a consequence, users can utilize efficient PIN interaction in trusted environments and spontaneously switch to the more secure *SwiPIN* interaction when needed. As both methods are based on *multiple temporal challenges*, interaction style can be switched even within single authentications. For example, a user could enter three digits using PIN and enter the fourth digit using *SwiPIN*. This allows very flexible adaptation.

If the secure interaction of a concept is not efficient, we recommend to design the concept in a way that allows users to adapt the usability-security trade-off to their current needs. We argue that this improves the overall usability of the system. In addition, we assume that extra costs for more secure interaction are easier to accept if they are linked to risky situations.

Study Design can Increase Error-rates and Systematic Analysis is Required

Unfamiliar secrets, limited training time and synthetic authentication tasks can increase error rates during user studies. The *XSide* project illustrated the importance of a qualitative error analysis and indicated that our taxonomy for gesture-based errors (see Section 3.3) is very useful for such tasks.

While error-rates were very high at first glance, the qualitative error analysis revealed that most errors occurred due to three reasons: mixing up sides, unintentional strokes and mirrored strokes. We argue that most of the errors were introduced by the specific user study design and would probably disappear in the wild. In the second user study, every fifth error happened as gestures were entered on the wrong (i.e., unintended) side. This error type does not exist in the wild as users can arbitrarily switch between the front and the back. However, we opted to control the sides in the study to compare the effects of front and back input. Similarly, unintentional strokes are rather triggered by programming issues than by interaction problems. We assume that most of the unintentional input can be avoided by applying better filtering techniques. Finally, mirrored gestures are most likely a consequence of using unfamiliar secrets. We assume that such errors will be significantly reduced when motor memory effects kick in [221]. Finally, as already discussed in Section 3.3, error rates might rise due to the fact that users deliberately fail to quickly restart the process. While the user study design of *SwiPIN* did not introduce such errors, we observed that performance data of the field study had limited value. The analysis of the mini questionnaires indicated that several input errors were logged while participants showed *SwiPIN* around.

We recommend to assess sources of errors qualitatively. Gesture-based input errors can be analyzed using the taxonomy provided in Section 3.3. Special attention needs to be paid to the study design as it may introduce input errors which are actually not possible in the wild.

Clever Interaction Design can Prevent Unwanted User Behavior

The systematic design and evaluation process revealed that the users' input characteristics influence the security of the respective system. First of all, both projects confirmed that efficiency can be a significant security factor. In several cases, traditionally entered PINs were observation resistant due to extremely fast interaction. In addition, we like to point out user behavior which was more specific to the tested system.

The security of *XSide* was mostly influenced by input speed and by the way participants lifted their fingers. If gestures are entered quickly and fingers are lifted not too much, it becomes very hard to distinguish single shapes. On the other hand, attackers could use longer breaks between the challenges as a good indication for the start of a new shape. In addition, the used side significantly influences the security of the system. As we were not able to test *XSide* in the field, the question remains if people would use side switching in a secure way. Interim solutions of *SwiPIN* indicated that security can be jeopardized by so called "bad lies" [73]. When participants had the freedom to select the starting point of their gesture, gestures were often performed directly on top of the intended digit. Such cues could be effectively used by attackers to guess the entered PIN. *SwiPIN inside* and *SwiPIN outside* illustrated that such behavior can be prevented by clever interaction design. Finally, physically separating output and input was shown to be the most effective solution. The *indirect interaction style* of *SwiPIN outside* naturally prevented "bad lies" without explicitly restricting the input.

We conclude that systematic evaluation strategies are required to identify unwanted user behavior. In many cases, unwanted user behavior can be prevented by adjusting specific design factors concerning the representation and the directness of interaction.

Field Tests are Important but Real-world Behavior is Hard to Assess

The importance of field evaluations was already mentioned. User acceptance, real-world behavior and social aspects can only be assessed in the wild. However, this Chapter showed several barriers which may limit the value or the feasibility of field studies.

Due to the lack of appropriate devices, we were not able to evaluate *XSide* in the wild. Therefore, it remains open work to answer the question if users will actually utilize side switching to protect their input from observation risks. Furthermore, the field evaluation of *SwiPIN* indicated general problems of such studies which were detected by applying an experience sampling method. Most of the logged *SwiPIN* events have not been triggered due to observation risks but due to other aspects. For example, participants played around or showed the system to someone else. This behavior significantly limits the value of the collected data as authentications were actually performed by several other persons. In addition, as observation risks are rarely perceived, the intended activation of secure interaction methods is rarely observed. As a consequence, we collected a very small number of *SwiPIN* authentications which did not allow parametric analysis. At the same time, we argue that this aspect is actually promising as it shows that *SwiPIN* was used as intended.

We conclude that adaptive interaction methods are rarely used in the wild. While this confirms that participants used the concepts in the intended way, it implies that field studies need to be performed over longer periods to draw valid conclusions. The use of experience sampling methods can be useful to distinguish conventional use from other events.

4.3.7 Limitations

Even though the concepts were systematically designed and thoroughly evaluated, there are inherent limitations which we address in this Section. First of all, the results are based on experiments with self-recruited samples. Overall, participants were younger and more tech-savvy than the general population. Thus, data might differ for other user groups and should not be generalized. Besides this general restriction, the projects had individual limitation which will be discussed below.

XSide

Even though the prototype was improved over the course of the project, the chunky form factor of the device may have introduced specific problems. Most notably, it was hard to interact with one hand. As a consequence, the observed performance data is based on two-handed interaction. It is likely that performance will improve with a slimmer device. However, this means that the results cannot be generalized to one-handed interaction. Another major limitation is based on the fact that we were not able to test the concept in the field over a longer period of time. Therefore, field behavior and memorability aspects are still unexplored. However, the results of the lab studies indicate that *XSide* gestures are easy to memorize. Finally, the security analysis might have been influenced by the fact that attackers knew the length of the passwords. For example, attackers might have observed only one stroke of a gesture but were aware of the fact that gestures were based on two strokes. Consequently, they had a chance of 33% to guess the right direction of the second (unobserved) stroke.

SwiPIN

The final concept of *SwiPIN* has been shown to be resistant against most observation attacks. Still, it was successfully attacked two times. This shows one limitation of the concept: *SwiPIN* allows attackers to reduce the guessing space by observing the input elements. Consequently, attackers can increase the probability of a correct guess from 0.01% to 0.16%. The problem could be solved by applying a full set of ten gestures. However, the pre-study indicated that using ten gestures would most likely reduce usability and user acceptance. Compared to *XSide*, another limitation of *SwiPIN* is that it requires visual attention and does not support motor memory effects. However, we assume that *SwiPIN* will only be used for a small number of authentications where this limitation is acceptable. Concerning the evaluation, the biggest limitations can be found in the field study. The performed study did not provide enough parametric data to draw valid conclusions about field performance. In addition, we did not collect field data on social aspects of the concept. For example, *SwiPIN* users might be afraid of communicating mistrust to other. A long-term study over several weeks needs to be performed to answer such questions. Finally, the results of the field study might be influenced by implementation issues as several participants reported that the deployed lock screen caused annoying errors. For example, some participants had to confirm the lock screen twice to unlock. Such technical issues might have triggered negative feelings which might have influenced the qualitative results.

4.3.8 Summary

In this Section, we presented two different concepts which allow efficient and observation-resistant authentication on mobile devices using simple touch gestures. Both concepts were designed following a systematic design approach using low-fidelity and high-fidelity prototypes. While the iterative process allowed stepwise improvements, we also learned from flawed interim solutions. Overall, this Section contributed by (1) presenting two feasible gesture-based concepts, by discussing the (2) results of several user studies and by revealing (3) general insights concerning the design and the evaluation of gesture-based authentication methods for mobile devices.

Both concepts utilized the same set of gestures to provide observation-resistance on demand (*RQ1*). *XSide* allowed to adjust the security by supporting eyes-free input. That is, users can adapt their input to the current context and authenticate out of sight of potential attackers while most authentications can be performed very efficiently on the front of the devices. *SwiPIN* achieved the same effect by allowing seamless integration into more efficient traditional authentication methods. Users would use PIN in most situations, but switch to dynamically guided gesture-input when more security is required. Efficient and secure dynamic guidance was realized by combining fixed input elements with randomized output elements. While first designs already indicated the feasibility of this approach, the overall performance was further improved by spatially separating input and output elements (*RQ2*).

The implementation of *XSide* required a novel type of gesture-based secrets. As eyes-free accurate pointing tasks were not feasible, we contributed by developing a self-contained gestural alphabet (*RQ3*). The alphabet uses very basic gestures (up, down, left, right) which can be combined to more complex shapes. The specific design of this password concept allows both efficient and effective eyes-free interaction and a large theoretical password space. In addition, *XSide* gestures are easily translated to textual strings and can therefore be handled by common infrastructures.

Finally, we learned how *directness* and *relation* influence observation-resistance and usability (*RQ4*). Allowing direct input jeopardized the security of *SwiPIN* as users tended to start their gestures on the intended digits. When we changed the interaction concept from direct to indirect input, both usability and security improved. In addition, even though *SwiPIN* could have been realized using self-contained gestures, target-orientation allowed to provide a smaller subset of gestures and improved efficiency.

4.4 On Increasing the Practical Password Space

In this Section, we will explore the design space in terms of gesture selection. The presented concepts aim at increasing the practical password space. This naturally implies that we focus on the gesture selection phase and not on the authentication event itself. Nevertheless, this Section will illustrate that the graphical design of gesture-based authentication mechanisms can be optimized in a way that nudges users to select a larger set of different secrets. Section 3.5 already revealed that current gesture-based approaches are prone to guessing attacks as gesture selection is very predictable. Analog to the approach presented in Section 3.5, the novel concepts will be evaluated in large-scale online studies and effects will be measured utilizing our novel similarity metric.

This Section will shed light on the following main *research questions*:

RQ1 Can background images serve as *guiding elements* to influence gesture selection?

RQ2 Can animations serve as *guiding elements* to influence gesture selection?

RQ3 How do users perceive such visual effects and how does it affect user acceptance?

RQ4 Which features of a grid-based gesture can be modified using such *static guidance*?

We present two different approaches. The first approach utilizes static background images and dynamic background images. The concept was evaluated in a lab study (n = 10) and an online study (n = 496). The second approach is based on dynamic foreground animations which are presented before the actual input takes place. The concept was evaluated in two online studies with 321 participants and 288 participants, respectively.

The results are based on detailed quantitative analyses of the collected gesture sets and on qualitative user feedback. We will learn that both background images (*RQ1*) and presentation effects (*RQ2*) can serve as effective guidance that nudges users to select a more diverse set of longer gestures. In this connection, the application of the similarity metric will show that overall diversity is increased when such effects are present. Most users were positive about the tested concepts and reported that the approaches simplified the gesture selection (*RQ3*). However, the results also indicate that such effects can be distracting and not all users can be influenced. Comparing the effectiveness of both approaches, we find that both concepts can increase gesture-length and the diversity of the set. However, even though presentation effects were shown to be effective in changing specific user behavior (e.g., starting position), we find that some habits are hard to break and the use of common shapes is hard to prevent.

Parts of this Section are based on a bachelor thesis [148] and a practical research project [149] by Anna Kienle. In addition, parts of this section have been included into von Zezschwitz et al., On Quantifying the Effective Password Space of Grid-based Unlock Gestures. To appear in Proceedings of MUM'16 [267]. Please refer to the beginning of this thesis for a detailed statement of collaboration.

4.4.1 Research Context and Motivation

Like traditional alphanumeric passwords (see Section 2.1), gesture-based secrets are often vulnerable to dictionary attacks as users select predictable shapes. In Section 3.5, we quantified the similarity of grid-based unlock gestures and found that using this widely accepted unlock method results in very similar and thus guessable secrets. In this regard, Section 2.3 discusses several countermeasures including password polices [197], password assignments [67, 89] and recommender systems [10, 241]. In conjunction with grid-based gestures, the use of strength meters has been suggested. Such concepts assess the strength of a given gesture and illustrate the expected guessing resistance using colors and progress bars [10, 229, 232, 241]. While the results indicate that such concepts can increase gesture length and complexity [229, 232], common selection strategies remain unchanged. For example, most users continue starting their gestures on the top-left corner of the grid [229]. Therefore, we assume that gesture similarity cannot be adequately addressed using common *feedback* mechanisms.

We argue that the design space of graphical gesture-based authentication mechanisms provides various options to influence selection behavior and gesture preferences using feed forward effects and *guidance*. While gesture length and gesture complexity are certainly important factors for guessability, this Section specifically aims to reduce the similarity of selected gestures by implicitly nudging users to modify their selection strategies (e.g., starting points). Psychological research indicates that such effects can be triggered with both subliminal [13, 150] and supraliminal stimuli [251]. Subliminal stimuli are provided over a very short period of time ($t < 50ms$) and are therefore not consciously perceived [154]. In contrast, supraliminal stimuli [25] are provided long enough to be consciously perceived. While both kinds of stimuli seem promising in general, we assume that subliminal effects are harder to provide on mobile devices. Firstly, hardware limitations may hinder the presentation of very short stimuli. Secondly, users may be distracted and may therefore not even focus the screen while a (very short) stimulus is provided. As a consequence, we opt for consciously perceivable stimuli which are more feasible for the mobile context. We specifically focus on the evaluation of background images and presentation effects.

Background images have already been proposed to improve the selection of Draw-a-Secret (DAS) [83]. The concept, called Background-Draw-a-Secret (BDAS) allows users to select personal background images. Even though BDAS was only evaluated in a low-fidelity paper prototyping study, the results were promising as users tended to select less symmetric shapes and started their input on different positions of the grid. Later, the positive effects were confirmed with a modified version of Pass-Go [202]. So called presentation effects [251] have been suggested in connection with Cued Click-Points [54], PassPoints [251] and Android unlock gestures [229]. Such systems provide supraliminal stimuli by presenting input elements in such a way that some elements are more focused than others. Thorpe et al. [251] suggested to influence the selection of PassPoints by presenting input elements in temporal order. Instead of showing the whole selection area at once, the system simulated a curtain which opens either from the left or from the right side of the screen. Siadati and Memon [229]

		Background	Presentation
Input	*Representation*	abstract	abstract
Elements	*Reusability*	limited	limited
Output	*Guidance*	static	static
Elements	*Feedback*	detailed	detailed
Interaction	*Relation*	target-oriented	target-oriented
Style	*Directness*	direct	direct
Element	*Temporal*	single	multiple
Arrangement	*Spatial*	random	random

Table 4.3: We apply specific static guidance to nudge users towards selecting a more diverse set of gestures. Presentation effects are applied before the input takes place while background images are presented during the whole process.

suggested to randomly highlight one specific cell of the 3*x*3 Android grid using blinking effects. Indeed, both projects revealed that users are nudged towards specific selections by applying such simple modifications of the user interface.

Based on the promising results of related work, we designed two concepts for grid-based unlock gestures. Concerning the design space, we especially investigate the effects of *output elements* and *temporal* as well as *spatial* arrangements. Both concepts have been evaluated in large-scale online studies to collect a sufficient amount of data that allows the application of the similarity metric presented in Section 3.5.

4.4.2 Threat Model

The threat model matches the situation illustrated in Section 3.5. An attacker gets in possession of a mobile device which is protected by a grid-based unlock gesture. He starts an educated guessing attack and tries to unlock the device within 20 attempts. From related work, he knows the most popular gestures and might even get additional information from smudge traces left on the screen. However, this time the owner of the device was nudged to use a different starting point which resulted in a less common gesture. As s consequence, the attacker is not able to authenticate within the limited number of tries and the owner's personal data remains secure.

4.4.3 Concept Overview

While two concepts have been evaluated in large-scale online studies, several other ideas have been rejected after preliminary evaluation. Table 4.3 categorizes the two main concepts using the design space for graphical gesture-based authentication mechanisms. However, this Section will also present concept ideas which have not been thoroughly evaluated.

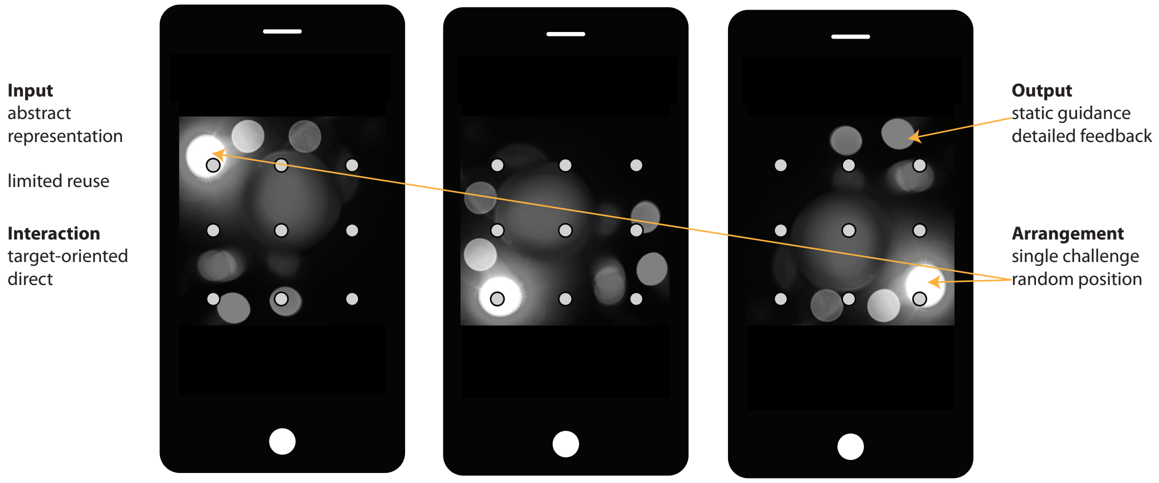


Figure 4.40: *Background* effects were applied using still images and animated images. An abstract image is randomly positioned in a way that specific input elements are highlighted.

Background Images

The concept is based on *static guidance*. For this purpose, a background image is displayed under the regular grid. The concept exploits the effect that users tend to select passwords based on visual hot spots [54]. Therefore, the background image is required to provide at least one visual hot spot which is used to highlight specific regions of the screen. We tested still images and animated images. The *spatial arrangement* of the image is adjusted to match one of the outer corners of the grid. A randomized spatial arrangement of the image helps to counterbalance the positions of the hot spots across different users.

Presentation Effects

Even though presentation effects utilize dynamic animations, we categorize them as *static guidance*. That is, the animations are performed independently from user interaction. The studied concept can be specified as providing multiple challenges as the effect is applied before the actual gesture selection takes place. For this purpose, the enrollment process starts with a grey square as indicated in Figure 4.41. The square conceals the grid of input elements. As soon as the gesture selection starts, the square disappears using a dynamic animation. The type of animation determines the presentation effect. We tested circular zoom-like effects, rotations and sliding effects similar to related work [251]. The spatial position of the focus area (target) should be counterbalanced to emphasize different input elements across users.

Further Concept Ideas

Besides the thoroughly evaluated concepts presented above, several other concept ideas came up during the ideation sessions.

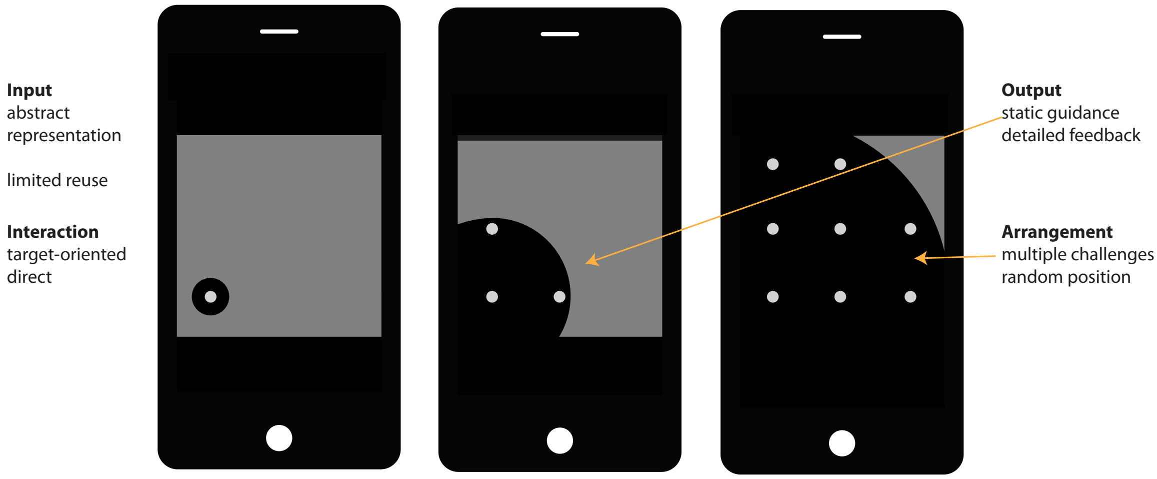


Figure 4.41: *Presentation* effects utilize dynamic animations. The animations are applied in a way that emphasizes specific input elements.

Personalizing Input Elements The idea is based on the assumption that users prefer specific shapes, colors and icons. Instead of using identical grid-cells, such concepts would use different representations for different cells. Over the course of the project, we discussed the utilization of symbols, abstract shapes and colors. A lab evaluation of an emoji-based prototype indicated promising effects.

Gamification and Rewards Applying gamification elements to gesture selection has been discussed in several dimensions. While obvious approaches would award secure gestures with desirable elements like stars and badges, more unusual solutions were also discussed. For example, the input area could comprise a hidden background image which would gradually appear when the respective regions are touched. The concept is based on the idea that curious users would touch more cells to see more of the background image's content.

Feed forward and Priming Subliminal and supraliminal priming effects were discussed. For example, secure gestures could be presented in advance of the selection tasks. Such gestures could either be part of an introduction page (supraliminal stimulus) or presented for a very short period of time (subliminal). The concept is based on the assumption that users would copy the presented gestures or adopt specific selection strategies (e.g., starting cells).

Dynamic Guidance In contrast to the evaluated concepts described above, dynamic guidance would respond to user interaction. The main approach is based on adaptive representations of input elements. While users enter a gesture, the representations of cells would change to prioritize more desirable cells for the next move. The concept could exploit several pop-out effects like color, size or distance. The approach is based on the assumption that users would accept the dynamic recommendations and select cells which are presented more prominent.

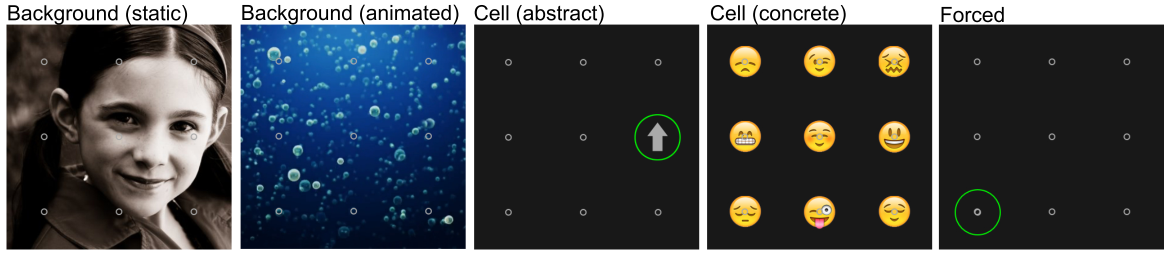


Figure 4.42: Five concept ideas were tested in a preliminary lab study.

4.4.4 Background Effects

In this Section, we aim at diversifying gesture selection by utilizing background images. Even though previous related work indicated promising effects for free-drawn shapes [83], the concept has never been evaluated for more restrictive grid-based unlock gestures. The first part of this Section outlines the evaluation strategy. The second part presents the results in terms of gesture selection and user feedback.

Prototyping and Evaluation Strategy

After an ideation phase and preliminary evaluations, we performed an online study using still images and animated images. This Section describes the prototypes and the study designs.

Evaluation 1: Preliminary Lab Study
First concept ideas were generated in an initial brainstorming session. There were no prerequisites for the ideas except that feasible concepts should visually highlight specific regions of the grid. The brainstorming resulted in five concept ideas which are illustrated in Figure 4.42.

***Prototype**:* The five candidate concepts were implemented as browser-based software prototypes using HTML5, CSS and JavaScript. *Background (static)* was based on a grayscale background image of a girl which contained several hotspots[11]. *Background (animated)* was implemented using a looped video[12]. The animated background illustrated bubbles floating from bottom to top. Besides potential effects on the starting cell, we were curious if the direction of the animation would influence the direction of gestures. In contrast to the background-based concepts, *Cell (abstract)* and *Cell (concrete)* were based on a modified representation of input elements. *Cell (abstract)* highlighted one node and indicated a specific direction using an upward pointing arrow. Instead of abstract cells, *Cell (concrete)*

[11] "love this face" by Jack Fussell, licensed under CC BY-NC-ND 2.0 (`https://www.flickr.com/photos/travelingtribe/3844008664`) – last accessed: 2016/06/06.

[12] "Air Bubbles Live Wallpaper", reproduced with permission by Eugene Pestov (`https://www.youtube.com/watch?v~=~fLbhOILIEcs`) - last accessed: 2016/06/06.

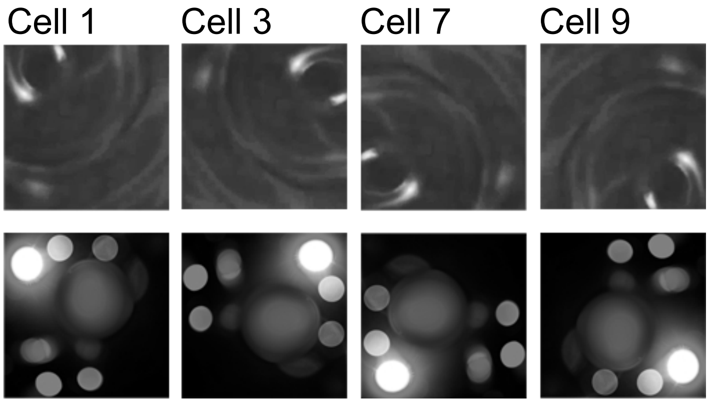


Figure 4.43: The final prototype presented *static* and *animated* images in different rotations, each highlighting one of the outer cells.

provided nine different emojis[13]. We assumed that individual preferences for specific emojis would trigger the selection of such cells. Finally, the *Forced* condition was based on a predefined a starting point and represented the baseline.

Design: The preliminary lab study was based on a repeated measure within participants design. The independent variable was *scheme* with five levels. The order of *scheme* was randomized. Data collection included selected gestures as well as qualitative user feedback.

Procedure and Setup: At the beginning, we informed the participant about the general goal of testing a novel authentication mechanism. User input was performed using a mouse and an office computer. For each concept, participants were asked to spontaneously select one unlock gesture. We did not provide further information on the scheme, nor did we set a selection policy. After each concept, we interviewed the participant and collected qualitative feedback. For example, we asked if there were any external factors which influenced the gesture selection. After all concepts were tested, participants answered a final questionnaire. Participants had the chance to win a 20 Euro voucher for an online shop.

Participants: We recruited ten participants with an age ranging from 25 to 34 years. Seven participants were female, three male. All participants reported to use at least one touch based mobile device on a daily basis. In addition, seven participants had gained experience in using Android unlock gestures.

Evaluation 2: Online User Study

Based on the results of the preliminary lab study, we selected the background concept to be evaluated in a large scale online user study. The online study aimed to collect an adequate number of gestures to allow investigating the concept's effects on starting points and on similarity.

[13] Based on Apple Color Emoji font

Hypotheses: The following main hypotheses were defined for the online study.

H1 *Background images* serve as *static guidance* and increase the diversity of user-selected grid-based gestures.

H2 Hot spots in *background images* serve as *static guidance* and nudge users to select specific target cells first.

H3 *Animated* images are as usable as *static* images but lead to stronger nudging effects.

Prototype: As a result of the lab evaluation, we opted to use more abstract background images for the final prototype. The background images were selected to have strong contrasts and clear hotspots. In addition, the animated condition was chosen to be calmer and therefore less distracting. The image used in the static background condition showed several spotlights on a dark background[14]. The animated condition was based on a looped video of a drop falling in water[15]. In contrast to the preliminary version, the concentric ripple did not indicate a certain drawing direction but specifically highlighted one cell. Therefore, the animated condition and the static condition were more comparable. To rule out the impact of colors, both concepts were based on grayscale images. Finally, both images allowed rotation in different directions while each rotation had the effect that the main hot spot matched one of the corner cells. To prevent uncontrolled side effects arising from the use of different interaction methods and device classes, the prototype was optimized for the use of laptops and desktop computers. Gestures were entered by pressing and releasing the left mouse button. Valid gestures had to comply with the common rules for Android unlock gestures. To illustrate the context of the enrollment, the grid was placed within an abstract representation of a mobile device. Interaction and user feedback was stored in a MySQL database.

Design: The online study was based on a between groups design with the two main variables: *Background (static)* and *Background (animation)*. For each concept, we tested four different rotation states. Therefore, participants were randomly assigned to one out of eight conditions. As illustrated in Figure 4.43, each of the conditions highlighted one of the cells "one", "three", "seven" or "nine".

Procedure and Setup: After the introduction page, the gesture input started. We requested two gestures per participant. The first gesture input was logged as training and allowed users to get familiar with the system. After the first input, a second input was performed using the same conditions. Please note that we did not provide any additional rules. That is, participants were free to either perform the same gesture twice or to perform two different gestures. Conditions were assigned based on a Round-robin scheduling. After the input task, participants were forwarded to a questionnaire to provide demographic data and to give feedback on the concept. Finally, all participants had the chance to win either one E-book reader or one out of five 10 Euro shopping vouchers.

[14] "Untitled" by I Love Trees is licensed under CC BY 2.0 (`https://www.flickr.com/photos/ilovetrees/2770624201` – last accessed 2016/06/06

[15] "Water ripples" by ChoiceSlides bought on `http://choiceslides.com/products/water-ripples` – last accessed 2016/06/06.

Participants: We recruited 503 participants using an university-wide mailing list. However, seven data sets had to be removed due to incomplete answers. The average age of the remaining 496 participants was 27 years (17-72). Almost half (51.2%) of the participants indicated to be female. Even though owning a mobile device was not a requirement for taking part in the study, 86% of the participants used a mobile device on a daily basis, 64% of them reported to use Android devices.

Results of the Gesture Analysis

In this Section, we assess the impact of background images on security. After summarizing the main insights gained from the preliminary lab study, the gestures of the online study are investigated in terms of composition aspects and similarity.

Evaluation 1: Preliminary Lab Study

The results of the preliminary lab evaluation were very promising as they indicated that all tested concepts have a strong effect on gesture selection. Both, background effects and modified cell layouts nudged users to start gestures at specific positions. In the Background (static) condition, participants tended to avoid crossing the girl's face which resulted in selecting the cells “two”, “three” or “nine” more often. The animated background image nudged users to perform gestures that followed the flow of the bubbles. That is, half of the participants started their gesture at the bottom of the grid and ended in the top row. Using the abstract cell modification, 90% of the participants started their gesture above of the arrow. In the concrete cell condition, participants tended to select happy emojis. Overall, we concluded that all tested conditions were well suited to modify selection behavior. As a consequence, the final decision had to be based on user feedback. As reported in the next Section, both background concepts were finally chosen for further evaluation.

Evaluation 2: Online User Study

The results are based on the analysis of 496 user-selected unlock gestures.

Basic Statistics: Overall, gestures were based on 6.08 (SD = 1.59) cells on average. A detailed analysis reveals that using the *static* prototype resulted in marginally longer gestures (6.18, SD = 1.61) than using the *animated* version (5.98, SD = 1.57). A one-way ANOVA comparing the average length of the collected gestures to the average length of the gestures collected under standard conditions (see Section 3.5) revealed a significant main effect, $F_{2,1000} = 61.30, p < .001$. Bonferroni-corrected post-hoc tests indicated that both background conditions led to significantly longer gestures than the baseline approach ($p < .001$). However, the comparison of the results of the animated condition and the static condition did not indicate significant differences ($p > .05$).

An analysis of starting positions revealed that many (36.1%) users still started their gesture at cell “one” and finished at cell “nine” (25.2%). While this indicates that most users still followed predictable selection strategies, we found that gesture complexity was increased. Compared to the baseline condition, participants utilized more special moves as 17.1% of the gestures included overlaps and 10.1% of the gestures comprised knight moves.

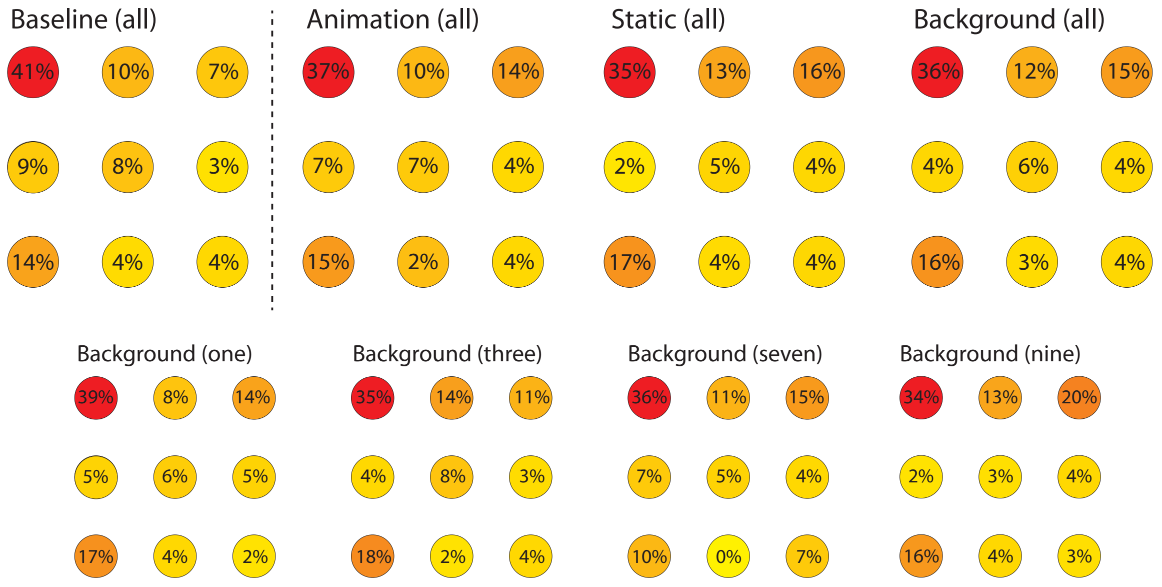


Figure 4.44: The analysis of starting cells indicates no significant impact of background images. However, the right side of the grid was used more often when background images were present.

Starting Cells: Figure 4.44 indicates that background images nudged users to use the right side more often. We ran a two-tailed binomial test to gather further insights. We defined the cell which was highlighted by a specific rotation as the *target cell*. Next, gesture selection was analyzed as binary event: Whenever the first cell of a selected gesture matched the *target cell*, the case was tagged as "success". All other combinations were tagged as "false".

The test calculated the probability with which the number of selected *target cells* (successful nudges) would have been the same within the baseline condition. That is, if no background image was provided. The comparison revealed only one significant association between the *static* background image focusing cell three and the actual selection of cell three ($p < .05$). All other comparisons revealed no significant effects ($p > .05$).

Popular Gestures: Overall, the -gesture was selected most often (n = 12). Table 4.4 illustrates the top five gesture groups for $n \leq 2$. The groups represent the most popular shapes. In addition, we illustrate the most popular single gesture in each group. The data indicates that both *animated* and *static* background images led to very similar shapes. Although different groups of participants contributed to each of the sets, the selected secrets show only minor differences in form and complexity. Two of the most popular gesture groups of the background study comprise gestures (,) which are also found in the top ranks of the baseline study (see Section 3.5) . In addition, the gesture " shows high similarity to two other popular groups of the baseline study (,).

While this indicates a close match of popular gestures in both baseline condition and background condition, the popular gestures differ in length and number. In the baseline study, all gestures of the top five groups were based on only four cells. In the background study,

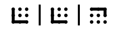
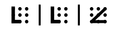
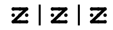
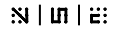
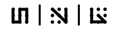
Rank	Top Gesture [O I S I A]	# Permutations [O I S I A]	# Occurrences [O I S I A]	% Dataset [O I S I A]
1		06 I 06 I 08	21 I 16 I 13	4.2 I 6.6 I 5.1
2		14 I 09 I 06	20 I 11 I 11	4.0 I 4.6 I 4.3
3		05 I 03 I 04	19 I 09 I 10	3.8 I 3.8 I 3.9
4		09 I 04 I 07	19 I 09 I 07	3.8 I 3.8 I 2.7
5		05 I 06 I 04	12 I 08 I 07	2.4 I 3.3 I 2.7

Table 4.4: The five largest groups for $n \leq 2$. For each condition, the table shows the most frequent (top) gesture of the group, the number of different gestures covered in each group, the accumulated absolute number of occurrences for all gestures in the group, and the covered ratio of the respective dataset. O≡overall, S≡static and A≡animated.

the overall top five of both background conditions comprises only one of such short gestures ( ). Two groups cluster length-five-gestures and the other two groups comprise gestures of length seven and nine, respectively. The fact that a gesture of the maximum length is found in the top five of the most popular gestures ( ) questions once again the importance of length as a security factor. According to the number and distribution of popular gestures, it should be emphasized that popular groups in the background condition cover less data than popular groups in the baseline condition. The baseline study revealed that the -gesture alone already covered over 11% of the whole data set. In contrast, none of the gestures selected with background images covers more than 4.2% of the whole data set. Assuming a similarity distance of $n \leq 2$, the top five gesture groups found in the baseline condition allowed to describe 29% of the data set. In the background study, the top five groups cover 18%.

Pattern Similarity: Analog to the analysis reported in Section 3.5, the collected gestures are analyzed using our similarity metric. Figure 4.45 illustrates the ratio of unique gestures for the distances $n \leq 2$. While the results indicate no significant differences for the type of background images, a comparison to the baseline data reveals that gestures of the background study are more diverse. Comparing the set of the baseline study with the set of the background study reveals that 31% of the gestures were duplicates ($n = 0$) when no background images were present. In the background study, 21% duplicates were chosen. By allowing more transformations, the similarity increases for all sets. However, more gestures in the background condition stay unique. By applying up to two simple transformations, the overall set of the background study shrinks to 58% of its size while the set in the baseline condition was reduced to 33%.

A length-dependent analysis of the gesture sets confirms the results of Section 3.5. Gestures which are based on four, five or seven cells are based on more similar shapes than gestures which are based on six or eight cells. Considering up to two transformations, we find that 33% of the length-four-gestures remain unique in the background condition. Without background images, only 17% stayed unique. In contrast, composing gestures of eight cells led to more diverse shapes in both conditions. Considering a distance of $n \leq 2$ as duplicates, reduced the set of background gestures to 79%, the baseline set to 60%.

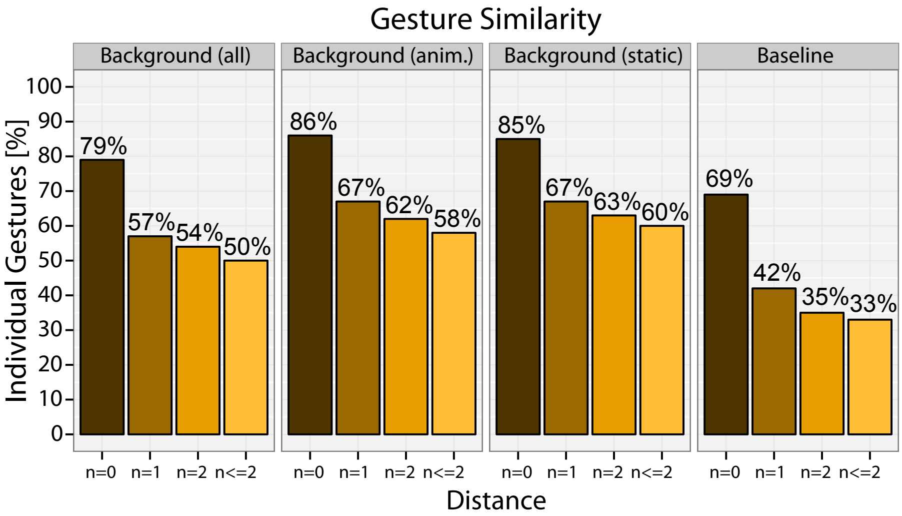


Figure 4.45: While both background schemes show similar distributions of similarity, selected gestures are overall more diverse than the baseline set.

User Feedback

In contrast to authentication, quantitative performance data has limited value for password selection tasks as such tasks are rarely performed. Therefore, we focus on the analysis of user feedback to asses usability and user acceptance.

Evaluation 1: Preliminary Lab Study

User feedback was the main factor for rejecting the *Cell (abstract)* concept and the *forced* condition. Both concepts were perceived as too restrictive and were thus unpopular. In contrast, most users were in favor of *Background (static)*. Even though two users suggested the use of more salient images. While the emoji-based concept was rated equally good, user feedback indicated that modifying the representations of input elements is more polarizing. One participant claimed that she disliked “smileys”, another participant claimed that the concept was distracting. Using *animated background* images led to most diverse reactions. While half of the users was in favor of the concept, the rest of the participants claimed that the animation was ‘too busy” and “made it hard to see the actual nodes”.

Overall, the results of the preliminary lab study indicated that using *static background* images and using *modified cells* were the most promising approaches. However, we were confident that the mentioned drawbacks of the *animated background* prototype could be minimized using a calmer animation. In addition, including animated backgrounds would enable a direct comparison of animated and static images. As we aimed to analyze if animated backgrounds are more effective than still background images, we selected both background schemes for further evaluation.

Evaluation 2: Online User Study
User feedback was collected using five-point Likert scales and open-ended questions. The open-ended questions were coded following the inductive coding approach reported in Section 3.2. Two coders analyzed answers concerning general feedback and the specific impact of background images. The process resulted in 362 and 318 code instances, respectively.

The five-point rating confirmed that *dynamic* background images are more eye-catching than *static* images. In the static condition, 54% of the participants agreed or fully agreed to have noticed the background image. In the dynamic condition, 68% indicated that the effect was consciously perceived. However, only the minority of participants would agree that the presented background images influenced their gesture selection. According to the Likert-based answers, this was the case for 6% of the participants in the *static* group and 5% of the participants in the *animated* group.

Most of the feedback focused on the general concept of grid-based gestures. However, the remaining 18% of the answers provided valuable insights into the effects of background images. Participants who reported positive effects often indicated an impact on starting points or mentioned visual details of the background image (n = 20). One user acknowledged that he *"[..] chose [the] starting point to be at the brightest spot in the image"*. Another participant reported that gesture selection was guided by *"The waves [which] were mostly in the upper left corner."*. In addition, participants indicated that the presence of background images facilitated gesture selection or made them rethink their selection strategies (n = 7). One participant stated that the *"big shining points left and right gave [her] an idea of the pattern [she] could use"*. Another user had *"[..] simulated the movement of the background"*.

Nevertheless, the feedback did also indicate negative effects and potential misconceptions. For example, the presence of background images led to reverse effects as one participant reported she opted for "the four dots where no white circles were [...]". Three participants specifically mentioned that they did not like the fact that the background image interfered with input elements of the grid. Five participants claimed that the background was confusing, distracting or irritating. While other users expressed general dissatisfaction with the visual design, others did not understand the purpose of the images. One participant stated she "*thought of [the image] as a plain background without further function*" and another user hypothesized that the image was provided to "*simulate the reflection of a real display*".

Summary

In this Section, we evaluated the feasibility of background images to implicitly nudge users to select a more diverse set of gestures. After an exploratory design phase and a preliminary lab evaluation, we evaluated two modifications of the background scheme in a large scale online study (n = 496). The collected gesture set was analyzed using traditional composition metrics (e.g., length) and the similarity metric presented in Section 3.5. While the lab-based evaluation of the concept was very promising, the results of the more realistic online study were mixed and indicated rather small effects.

On the one hand, we found that the selected gesture set of the online study was indeed more diverse than the set collected under baseline conditions. Overall, users utilized more cells and the practical gesture space remained more diverse. Therefore, we accept the hypothesis that background images diversify the selection behavior (*H1*). However, since a majority of the participants reported that the background images were not recognized, the effects seem to be subconsciously perceived. On the other hand, we cannot accept hypothesis *H2*. While some users reported that visual hotspots influenced their starting positions, the statistical analysis indicated only one significant association. Furthermore, we found no significant differences between *animated* background images and *static* background images and reject *H3*. Even though users stated that animations were more eye-catching, they had the same effects on gesture selection as static images.

Overall, we conclude that background images can have positive effects on gesture selection but the impact is rather small. The next Section will investigate the effects of foreground presentation effects and show that such concepts have stronger impacts on selection behavior. This is especially true for the selection of starting points.

4.4.5 Presentation Effects

This Section investigates the feasibility of foreground presentation effects [251] in connection with grid-based gestures. In contrast to background images, such effects are triggered by simple time-dependent animations which unfold the input area step by step before the input takes place. Therefore, presentation effects are not present during gesture selection. The first part of this Section outlines the evaluation strategy. The second part presents the results of the gesture analysis and the user feedback.

Prototyping and Evaluation Strategy

The evaluation of background effects indicated that (repeated-measures) lab studies have limited value when dealing with gesture selection tasks. As a consequence, we decided against lab evaluations. That is, after a qualitative focus group discussion and after a preliminary online study, we performed a large-scale evaluation analog to Section 4.4.4.

Evaluation 1: Preliminary Online User Study

An initial brainstorming generated several ideas how to implicitly influence gesture selection. As already pointed out in Section 4.2.3, we considered visual modifications of the interface, gamification approaches, feed forward mechanisms and dynamic feedback. The different approaches were then discussed in a focus group with five smartphone users. Participants of the focus group rated the concept ideas using ten-point scales. The results indicated that presentation effects were widely accepted and overall most promising. To find a set of feasible animation effects and to get a basic idea of animation speed and animation style, we then built dynamic mock-ups using Adobe Flash. Informal tests resulted in a set of three animations which were evaluated in a preliminary online study (see Figure 4.46):

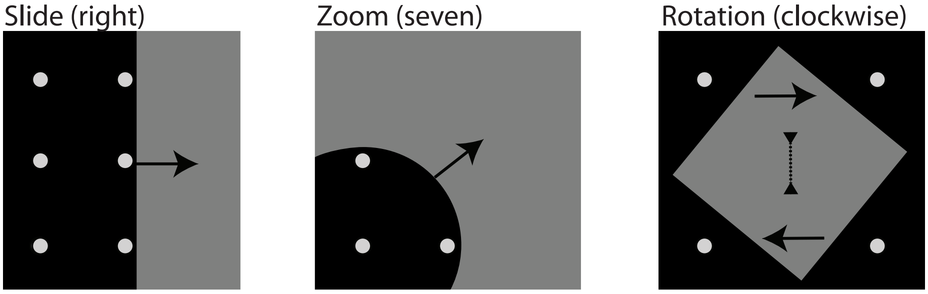


Figure 4.46: Three presentation effects were evaluated in the preliminary online study. The zoom effect was most promising and therefore tested in the final user study.

Rotation is based on a square which rotates and shrinks at the same time. The animation starts as an overlay of the input field and disappears in the center of the grid. The primary goal of the animation is to influence the drawing direction.

Slide is based on a square which slides out of the grid. The animation starts on the grid and gradually disappears by sliding along the x-axis or along the y-axis. The animation reveals the rows or the columns step by step and aims to influence the starting region.

Zoom is based on a circle which is gradually increased. The effect starts with an overlay which then disappears by opening a hole. The animation aims to influence the starting position and therefore starts on a specific cell of the grid.

***Prototype**:* We built a responsive web-based prototype using PHP, JavaScript and HTML5. The animations were implemented using CreateJS[16]. The prototype was based on the software used in the background image study and mimicked the standard Android interface. In contrast to the previous study, the prototype was optimized for mobile devices. The use of mobile devices was checked using MobileDetect[17]. All effects started as a grey overlay and disappeared according to the respective animation.

***Design**:* The preliminary online study was based on a between-groups design. The independent variable was *scheme* with three levels: *Slide*, *Rotation* and *Zoom*. *Slide* was tested in each of the four directions: west, north, east and south. *Rotation* was tested in anti-clockwise rotation and clockwise rotation. *Zooms* originated from the cells “one”, “three”, “five”, “seven” and “nine”. In addition, we tested a baseline condition where the square disappeared without animation. The twelve conditions were randomly assigned. We logged user interaction and collected qualitative user feedback using a short questionnaire.

***Procedure and Setup**:* After an introduction, users were forwarded to the input task. The input task requested two gestures while the first gesture was logged as training input. Both tasks started with a grey overlay as indicated in Figure 4.41. The square disappeared as soon as the participant pressed a start button. In the first input task, the square disappeared without

[16] `http://createjs.com` – last accessed: 06/14/2016.

[17] `http://mobiledetect.net` – last accessed: 06/14/2016.

animation. In the second input task, one of the twelve conditions applied. That is, the square disappeared either without effect (baseline) or with a predefined presentation effect. In both task, the input had to be confirmed by reproducing the gesture once. Finally, participants provided feedback using a short questionnaire.

Participants: 321 participants completed the study. The average age was 25 years (SD = 5.2). 165 were female, 156 were male. The majority (64%) of the sample used Android smartphones, 58% of them indicated to use unlock gestures.

Evaluation 2: Online User Study

The preliminary online study showed no significant differences between the conditions but indicated that the *zoom* out effect was most promising. Therefore, the online study was repeated focusing on the four conditions of *zoom*. The main goal was to collect enough samples per condition to allow a valid assessment of nudging effects.

Hypotheses: The following main hypotheses were defined for the online study.

- *H1* *Presentation effects* serve as guidance and increase the diversity of user-selected grid-based gestures.
- *H2* *Zooming* out of a specific cell nudges users to start their gestures on the respective cell.
- *H3* *Presentation effects* are more effective than background images as the guidance is more specific.

Prototype: We used the prototype of the preliminary online study. However, since the first evaluation had revealed several problems concerning the presentation of animations using mobile browsers, the software was optimized for desktop computers. Therefore, MobileDetect was configured to prevent the use of mobile devices.

Design: The study was again based on a between-groups design. *Zoom* was tested in four versions. Therefore, the independent variable was *target* with four levels: *one*, *three*, *seven* and *nine*. Participants were randomly assigned to one of the four conditions.

Procedure and Setup: The procedure was the same as described in the preliminary online study. The only modification was based on the fact that participants used a desktop computer and a mouse to enter the gesture. Input was performed by pressing the left mouse button.

Participants: We recruited 292 participants using a university-wide email list. Four data sets had to be removed due to incomplete answers. The average age was 26 years (SD = 8.9). 167 participants were female, 121 were male. Most (78%) participants used Android smartphones on a daily basis, 60% of them indicated to use unlock gestures.

Results of the Gesture Analysis

In this Section, the collected gestures are investigated in terms of composition aspects and similarity aspects. Even though we focus on the results of the main study, we begin by presenting the main insights of the preliminary online study.

Evaluation 1: Preliminary Online User Study

Overall, the preliminary online study indicated only minor effects. Statistical analyses of composition aspects revealed no significant differences between the conditions. In addition, a manual inspection of the gestures indicated no effects on the sense of rotation. However, the results indicated a minor impact on starting positions. For example, cell "three" was used by 3% of the participants in the baseline condition and when a *slide* animation was focusing the left side of the grid. In addition, when a *zoom* originated from cell "one", no participant started the gesture on cell "three". In contrast, animations which focused the right side of the grid tended to nudge user to start gestures on the right side. 10% started their gesture on cell "three" when the region was emphasized by a *slide* animation and 24% started their gesture on cell "three" when it was highlighted by a *zoom* effect. Similar effects were indicated for the cells "one" and "seven". Overall, the *zoom* effect seemed to have the strongest impact on starting points and was therefore selected for further investigation.

Evaluation 2: Online User Study

Analog to the background user study, the data was analyzed in terms of popular gestures, similarity and starting cells. Due to an error in the assignment algorithm, group sizes were not perfectly counterbalanced and ranged between 54 and 113 samples per condition. Nevertheless, we argue that such sample sizes meet the requirements for statistical analysis.

Basic Statistics: The overall length was 5.95 (SD = 1.61) cells on average. In addition, gesture length rarely changed between the different *zoom* conditions. We measured a minimum of 5.6 cells in group "three" and a maximum of 6.1 cells in group "nine". We performed a one-way ANOVA to compare the average length of the collected gestures to the average length of gestures collected in both the baseline study and the background study. The results indicate a significant main effect, $F_{2,1290} = 65.89, p < .001$. Bonferroni-corrected post-hoc tests confirm that users selected significantly more cells when presentation effects were present. However, compared to the background study, no significant differences were found.

Overall, 41.3% of the participants started their gesture at cell "one" and 33.3% of the gestures ended with cell "nine". These numbers match the observations in both the baseline study and the background study. However, participants of the presentation effect study used more special moves as 15.6% of the gestures comprised overlaps and 9.0% included knight moves.

Starting Cells: Figure 4.47 illustrates the ratio of selected starting cells in different conditions. Comparing the overall data indicates no significant effects. However, a comparison of the different conditions indicates significant influences. The strongest effects can be found for cell "three" and cell "seven". When cell "three" was focused, 27% started their gesture on the target cell. When cell "seven" was emphasized, the target cell became the most often (42%) selected starting point.

A two-tailed binomial test confirms that foreground presentation effects significantly influence starting cells. A comparison to the baseline data indicates significant differences for the cell "three", "seven" and "nine". The respective cells were selected more often than

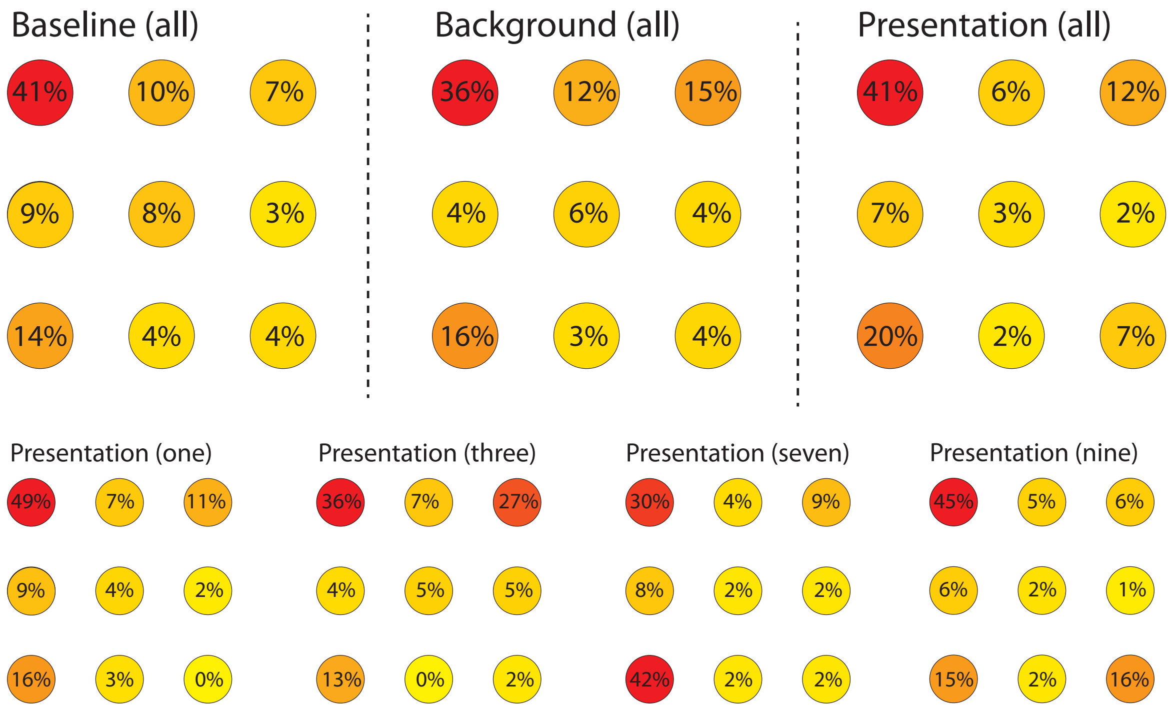


Figure 4.47: The analysis of starting cells indicates a significant impact of presentation effects. This was especially the case, when cells "three" and "seven" were focused.

expected under baseline conditions (all $p < .001$). However, the use of cell "one" was not significantly different. A comparison to the background study reveals the same differences. Cell "three" ($p < .05$), cell "seven" ($p < .001$) and cell "nine" ($p < .001$) were selected significantly more often when presentation effects were applied.

Popular Gestures*:* "1235789" () was selected most often (2.8%). This matches the data of the background study, where 2.4% of the users selected this specific gesture. Table 4.5 illustrates the most frequently selected gestures of popular gesture groups for $n \leq 2$.

While most popular gestures are based on common shapes, we find interesting differences compared to previous studies. Compared to the baseline condition, participants selected more cells. Overall, three of the most frequently selected gestures were based on five cells, the other two groups are based on six and seven cells, respectively. In addition, the analysis of the individual gesture sets reveals that users tended to base their gestures on more complex shapes. We found several gestures which were not present in the other two user studies. Most remarkably, two popular gestures comprised overlaps (,). In addition, popular gestures actually started on the target cells. This was the case for the gestures and when cell "three" was focused and for the gestures , and when cell "seven" was focused. Finally, emphasizing cell "nine" resulted in the use of the -gesture which is based on a common shape but started at cell "nine".

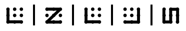
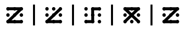
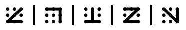
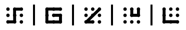
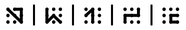
Rank	Top Gesture [A I 1 I 3 I 7 I 9]	# Permutations [A I 1 I 3 I 7 I 9]	# Occurrences [A I 1 I 3 I 7 I 9]	% Dataset [A I 1 I 3 I 7 I 9]
1		06 I 03 I 02 I 03 I 03	16 I 06 I 05 I 04 I 06	5.6 I 8.6 I 9.1 I 7.5 I 5.4
2		03 I 04 I 03 I 03 I 02	15 I 04 I 03 I 03 I 05	5.2 I 5.7 I 5.5 I 5.7 I 4.5
3		07 I 02 I 02 I 01 I 04	10 I 03 I 02 I 02 I 05	3.5 I 4.3 I 3.6 I 3.7 I 4.5
4		05 I 02 I 02 I 02 I 02	10 I 03 I 02 I 02 I 04	3.5 I 4.3 I 3.6 I 3.8 I 3.6
5		05 I 01 I 02 I 02 I 03	09 I 02 I 02 I 02 I 04	3.1 I 2.9 I 3.6 I 3.8 I 3.6

Table 4.5: The five largest groups for $n \leq 2$. The table shows the most frequent (top) gestures of a group, the number of different gestures covered in each group, the accumulated absolute number of occurrences for all gestures in the group, and the number of occurrences as a ratio of the respective dataset. We report the whole set ($\equiv$ A) and the individual conditions.

Even though the most popular gestures are based on common shapes, no group covers more than 5.6% of the overall gesture set. Nevertheless, the top five groups describe 20.8% of the selected gestures. Within the same conditions, popular groups tend to describe slightly more gestures as the percentage coverage ranges between 22% and 26%. For example, the -shape covers 9.1% of all gestures selected in condition "three".

***Pattern Similarity**:* Figure 4.48 illustrates the ratio of unique gestures after applying $n \leq 2$ transformations. Overall, the similarity values match the values observed in the background images study. That is, users selected a more diverse set of gestures than in the baseline study.

Overall, 23% of the gestures were duplicates ($n = 0$) and the amount of unique gestures drops to 48% when considering $n \leq 2$ transformations. The chart on the right illustrates the similarity values after excluding all cases where participants were nudged to use cell "one". Indeed, the results indicate that the selected gestures are more diverse when participants were nudged towards cell "three", cell "seven" and cell "nine". An analysis of the individual gesture sets confirms this assumption. Focusing on duplicates reveals that 14% of the gestures were selected at least twice when users were nudged towards cell "one". Emphasizing cell "three" or cell "seven" resulted in 5% duplicates and 4% duplicates, respectively. When cell "nine" was focused, 10% of the patterns were selected at least twice. Considering up to two transformations, the amount of unique gestures drops to 66% ("one"), 78% ("three"), 75% ("seven") and 64% ("nine"), respectively.

The length-dependent analysis confirms the findings of previous evaluations. Considering duplicates, length-five (28%), length-seven (30%) and length-nine (25%) gesture show the highest values. In contrast, the set of length-six (18%) and length-eight (6%) gestures comprised less duplicates. In comparison to previous user studies, length-four gestures also comprised less duplicates (16%). Considering up to two transformations indicates three similarity classes. Gesture sets with four and five cells are reduced to 43% and 40% of the original size. Gestures with six, seven and nine cells are reduced to 53%. Gestures which are based on eight cells remain most diverse (75%).

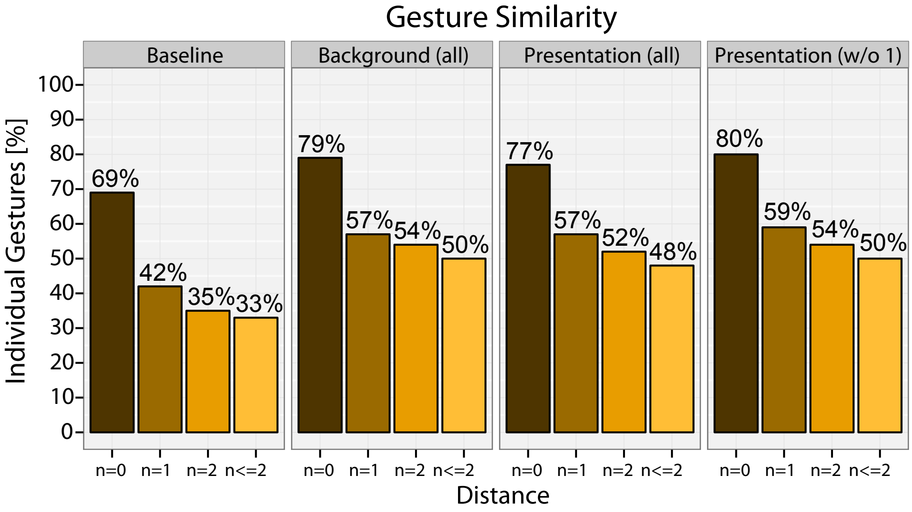


Figure 4.48: Compared to the baseline set, presentation effects led to a more diverse gesture choice. This is especially true, when the cells "three", "seven" and "nine" were focused.

User Feedback

The user feedback provides valuable insights into users' mental models and into the specific impact of the tested schemes. We focus on a qualitative analysis of user statements.

Evaluation 1: Preliminary Online Study

The feedback was voluntary and collected with open-ended questions. About half of the participants (36% overall) who answered the question reported that their gesture selection was not affected by the animations. Many participants claimed that they already "had a pattern [..] in mind before [..]". However, some participants stated that the effects nudged them to "be different" or to select a "more complicated password". One participant reported that "the displayed moving shape intuitively created a pattern in [her] mind." Nevertheless, it became apparent that a lot of participants did not recognize the animations. Comments like "which behavior?" or "did I miss something?" indicated that animations may not have been correctly displayed on some devices. Later tests confirmed the assumption. As a consequence, the final user study was performed using desktop computers.

Evaluation 2: Online User Study

The feedback of the online study was based on two open-ended questions. The answers were analyzed using an inductive coding approach. With the first question, we investigated the general impact on the gesture choice. Most participants reported general usability aspects. Overall, 20% of the gestures were selected as they were "easy to enter" and 12% of the participants claimed their gesture was "easy to remember". For example, one participant stated: "I think the U's are easy to remember". 14% specifically named "security" as one reason for their gesture choice. Interestingly, only 7% said that the gesture was "fast to enter". Other reasons included eyes-free interaction (1%), playfulness (1%) and convenience (1%). Only four participants named the presentation effects at this point.

Next, we specifically asked for the impact of the animations. 42% claimed that the animation had no impact on their gesture selection: "In no way, because I chose a pattern beforehand". However, 11% specifically mentioned that the effects influenced their starting point. Some (2%) participants reported indirect influences as they "decided against the [emphasized] starting point [..]". Individual statements indicated also misconception: "The starting point was influenced, but there was no other starting point possible, was it?". Finally, 8% of the participants did not (consciously) notice any effect.

Summary

We investigated the feasibility of presentation effects to influence grid-based gesture selection. After a first ideation phase, three different animations were evaluated in a preliminary online study (n = 321). The results indicated that circular zoom-like animations were most promising. A second online evaluation (n = 292) tested the presentation effect in detail. The collected gestures were analyzed analog to the approach presented in Section 3.5. Overall, the selected gestures were more diverse and more complex than gestures selected in the baseline study. In addition, the results indicate that foreground presentation effects have a strong impact on starting cells.

The analysis revealed that participants selected a diverse set of complex gestures. This was especially the case, when animation effects focused on uncommon starting cells like "three" and "seven". We therefore accept *H1* and argue that presentation effects can increase the diversity of user-selected gestures. Furthermore, we accept *H2* since emphasizing specific cells effectively nudged users to start their gestures on such cells. In condition "seven", cell "seven" actually became the most used starting cell. Focusing on similarity and traditional composition metrics (e.g., length), presentation effects seem as effective as background images. However, when focusing on the impact on starting cells, presentation effects seem more effective than background images. We therefore accept *H3* but note that the advantage is specifically related to starting positions.

We conclude that presentation effects are indeed suited to diversify the selection of starting points in a large data set. However, the concept cannot guarantee the selection of specific cells since a majority of users indicated that they are not affected by such an approach.

4.4.6 Discussion and Implications

We gained valuable insights into the impact of presentation effects and background images on gesture selection. This Section links the main results and discusses their implications.

Graphical Guidance can Effectively Influence Gesture Selection

Both background images and presentation effects influenced gesture selection. We found that users selected more cells than in the baseline user study and gesture sets were overall more diverse. This indicates that simple graphical modifications can effectively influence gesture selection. In contrast to traditional measures like password policies and strength meters, the presented cues affected gesture selection more implicitly. Even though we do not claim that the presented concepts are more effective than traditional measures, it is important to understand that such small changes of the user interface can have significant effects. In terms of effectiveness, we conclude that presentation effects are better suited than background images. Firstly, users were less distracted as the effects disappeared before the input took place. Secondly, presentation effects had a significant impact on the starting position. Overall, we conclude that well designed guiding elements have the potential to change selection behavior. In this connection, well designed guidance is not intrusive but eye-catching and communicates a simple message (i.e., "use this cell").

Gesture Selection is Affected in Various Ways - Knowingly and Unknowingly

The user feedback indicated that background images and presentation effects can affect gesture selection in various ways. The inductive coding of the statements revealed positive and negative effects. While some effects were intended (e.g., selecting the target cell first), other effects were not considered at first. For example, participants recognized the cues, reconsidered their selection and deliberately selected cells on the opposite side of the target cell. Other participants accepted the recommendation without reconsideration, partly believing they were forced to select the target. On the other hand, the feedback indicated that implicit guidance can generally simplify gesture selection as some users stated that the cue gave them a general idea of what gesture to take. Finally, we assume that some participants who claimed that they did not perceive any effects were subconsciously influenced. We conclude that presentation effects and background images can have a broader range of efficacy than traditional feedback mechanisms (e.g., strength indicators). However, this also holds true for negative effects as some users rated the concepts to be distracting and confusing.

Habits are Hard to Break and Popular Shapes are Hard to Prevent

Overall, the analysis indicated that selection habits can indeed be changed with background images and presentation effects. Most notably, the presentation effects were shown to modify starting positions of gestures. However, we found that such effects are more likely if cells are more in line with common starting positions and the directions of reading. And as a consequence, nudging effects became weaker when targets were uncommon. Moreover, participants still favored common shapes even if they were nudged to using uncommon starting position. For example, the ⁝⊏-gesture was among the most popular groups when cell "nine" was targeted.

One reason for this might be that many users actually used grid-based gestures on a daily basis and therefore were already heavily biased towards specific selection strategies. We assume that unconstrained guidance is not suited to entirely change existing habits. We thus conclude that suitable concepts should not try to change existing behavior completely, but can find useful opportunities in aiming at slight changes within these general habits.

Some Users are Resistant to Presentation Effects and Background Images

A large portion of the participants assigned to the background study and about 35% of the participants assigned to the presentation effect study stated to not have noticed the effects. Furthermore, only a minority of the participants agreed that effects did actually influence their gesture choice. However, the results showed that users in both studies selected a more diverse set of longer gestures. This indicates that several participants have unconsciously changed their selection strategies. Nevertheless, we expect that such concepts affect only a subset of users. We assume that the characteristics of both the used background image and the used animations significantly influence its perceptibility. In addition, the response to different presentation types might vary between users. We therefore conclude that the interplay between specific features of the visualization and the impact on particular user groups needs further investigation.

Evaluation Strategies Matter and Lab Studies seem Inappropriate

During the development process, we evaluated the background schemes and three other designs in a lab-based user study. The outcome was very promising as all concepts had a significant impact on the chosen starting points. In addition, this impact was confirmed by the participants' qualitative feedback. In contrast to such promising results, the online studies indicated only minor effects, especially when background images were applied. Interestingly, previous evaluations [83, 251] which also indicated significant behavior changes were also performed in the lab. Even though this previous work considered different authentication schemes which might lead to different results, we argue that one reason for the strong effects might be the evaluation strategy. Since our lab study was designed following a repeated-measures design, participants were exposed to different designs and adapted their behavior accordingly. In the online study, each participant was exposed to only one condition. Our results indicate that the study type has a significant impact on the outcome and repeated-measures lab studies are not feasible to gather valuable insights concerning behavior changes. We argue that in such cases, data needs to be collected outside of the lab and participants should be assigned to only one condition.

4.4.7 Limitations

Considering the evaluation strategies, not all confounding factors could be ruled out. The baseline dataset was collected via Amazon Mechanical Turk, mostly completed by US citizens. In contrast, the background image study and the presentation effects study were both conducted in Europe. This could have had an impact on specific aspects of the gesture selection. In addition, the system setup was slightly different as gestures were entered using a mouse and a desktop computer. Finally, it might be possible that some of the observed differences resulted from testing different groups and not from applying different effects. However, since we invited a large number of participants and all participants received the same instructions, we believe that the data is comparable across different studies. This assumption is supported by the fact that important selection strategies could be observed across the different user studies. Finally, it is important to note that participants of all groups were younger and higher educated than the general population. Therefore, the results might not be representative of other age groups, cultures and contexts.

Even though the used background images and animation effects were selected based a thorough design process including preliminary evaluations, it is possible that other visualizations trigger different user behavior and might actually have stronger effects. Nevertheless, the presented concepts contribute to the understanding of such systems and illustrate how visual cues can be utilized in a way that influences password selection behavior.

4.4.8 Summary

In this Section, we presented several concepts which utilized output guiding elements to implicitly influence gesture selection. The design and evaluation process was based on low-fidelity mock-ups and high-fidelity prototypes and comprised both lab studies and online studies. In summary, this Section contributed by (1) presenting and investigating two concepts which utilized background images and foreground animation effects. The results indicated that both concepts are feasible (2) even though users' habits were hard to change. In addition, we discussed general insights (3) into selection behavior, gesture similarity and the evaluation process. Finally, we confirmed the utility (4) of the similarity metric which was proposed in Section 3.5.

Concerning our research questions, we conclude that background images can serve as *static guidance* and indeed increase gesture diversity. However, common selection behavior like preferences for specific starting positions are hardly influenced (*RQ1*). For this purpose, dynamic animations seem overall better suited (*RQ2*). The results indicated that presentation effects are overall more effective and less distracting. Therefore, we would generally recommend this type of effect. However, background images might be an option when dynamic animations are not feasible (e.g., due to hardware limitations). Both concepts were well accepted (*RQ3*) and users reported mostly positive effects. For example, the visual cues made them reconsider the gesture selection or inspired them to use a different shape. However, especially background images were also rated confusing and sometimes triggered negative feelings. Finally, we found that not all composition features can be addressed equally well (*RQ4*). While gesture complexity was increased and starting positions could be slightly changed, most participants still used a limited set of common shapes.

4.5 Result Aggregation

We presented various research projects which investigated how mobile authentication methods can be designed in a way that improves both usability and security. Motivated by the findings presented in Chapter 2 and Chapter 3, we addressed three major problems of current gesture-based authentication: Smudge attacks, observation attacks and guessability. Overall, ten concepts were evaluated in detail using low-fidelity prototypes and high-fidelity prototypes. Each candidate concept was iteratively improved and tested in lab or field studies.

Based on the findings presented in Chapter 3, all concepts aimed at providing the same usability as current solutions while improving practical security. Overall, the presented concepts showed low error rates and supported very efficient interaction styles. For example, we learned that efficient input can even be supported by randomized *spatial arrangements* (Section 4.2), even though the *temporal arrangement* plays a significant role for perception. In addition, we learned how the *directness* and *relation* of gestures can be configured to design observation-resistant authentication mechanisms which can provide security on demand (Section 4.3). *SwiPIN* showed that target-orientation can be useful to prevent undesired user behavior while *XSide* illustrated that self-contained gestures enable eyes-free interaction.

Output elements were especially utilized to guide gesture input. *SwiPIN* and *Connect Four* used such elements to prevent observation attacks and smudge attacks. Section 4.4 illustrated that background images and presentation effects can be used to guide gesture selection. At the same time, *XSide* illustrated that output elements are optional and gesture-based authentication can completely forgo feedback and guidance. Even though Chapter 3 indicated that memorability is a minor problem in the mobile context, we identified various influencing factors. For example, the design of *Marbles* and *Marbles Story* illustrated how memorability is affected by the used *representation* of input elements. In addition, we learned how observation-resistant concepts can be designed in a way that supports motor memory effects.

Finally, the projects revealed that gesture-based interaction and graphical interface design can help to improve established mechanisms. For example, we learned how gestures can help to protect PIN-entry from observation attacks and how graphical design can help to increase the practical password space of Android unlock gestures. On the other hand, we showed how the design space can be exploited to design completely novel authentication concepts which solve the same problems. For example, the evaluation of *Marbles* indicated that secrets are less predictable when input elements are randomized and selection is based on *multiple temporal challenges*.

In summary, we conclude that all factors of the design space play a significant role for the design of usable and secure authentication mechanisms. Even though the aspects are interconnected in various ways and slightly different design decisions may lead to completely different outcome, the results indicate main dependencies. The next Chapter will structure the findings of both the design space and the problem space and derive recommendations for action. In addition, we will provide valuable insights into the evaluation of mobile authentication concepts.

III

Reflections & Conclusion

Chapter 5

Implications of the Interconnection of Design Space and Problem Space

The whole is greater than the sum of its parts.

– Aristotle, philosopher (384 BC - 322 BC) –

Chapter 3 provided valuable insights into the current state of mobile authentication. In addition, Chapter 4 illustrated how graphical gesture-based authentication can help to prevent smudge attacks, observation attacks and guessing attacks. In this Chapter, we combine the insights of both basic research and design-oriented research and illustrate the interconnection of design space and problem space.

Firstly, Section 5.1 revisits the problem space and presents ten design objectives for usable and secure authentication on mobile devices. Secondly, Section 5.2 provides assistance for a goal-oriented design and development process. We outline the impact of single design factors and illustrate how design decisions can affect the performance of the resulting authentication mechanism. In addition, we provide recommendations for a systematic development process and discuss how the gained knowledge facilitates future developments. Finally, Section 5.3 addresses important aspects of the evaluation of mobile authentication mechanisms.

5.1 Requirements and Design Objectives for Mobile Authentication

In the first part of this Section, we revisit the problem space and discuss the main factors in the light of the results of this thesis. In the second part, we present concrete design objectives for feasible gesture-based authentication on mobile devices. While this Section already mentions some influencing design factors and names recommended evaluation strategies, Section 5.2 and Section 5.3 provide detailed recommendations for the design and the evaluation of mobile authentication mechanisms.

5.1.1 Revisiting the Problem Space

Section 3.1.1 provided a definition of the problem space of graphical gesture-based authentication on mobile devices. This Section revisits the main factors and discusses new insights. For the sake of brevity, security aspects will be combined to password exposure and password space. Accessibility and divulgation resistance will not be addressed as both factors were out of scope of this work and were therefore not systematically evaluated.

Effectiveness

Effectiveness is the most important usability factor. If users are not able to authenticate, the concept is not feasible. The results confirmed that the quantitative assessment of error rates is a useful approach to assess effectiveness. However, we also found that the analysis should not be limited to counting errors as the results indicated that the type of error recovery might have a bigger impact on the perceived effectiveness of a system than the number of errors. We found that critical errors or memorability-related errors are rather seldom in the mobile context and most errors are based on slips. Since such slips frequently occur, recovery from errors is more important than error prevention. This was confirmed by the fact that users often deliberately failed in order to start over quickly. In addition, we found that error rates are artificially increased by study procedures which request the use of unfamiliar secrets or introduce error types which are not even possible in the wild. Overall, this renders the sources of errors more important than the number of errors and indicates that effectiveness needs to be assessed quantitatively and qualitatively. Section 3.3 presented a taxonomy for gesture-based input errors which can help to gather detailed insights into the effectiveness of a given (gesture-based) concept and indicates potential usability problems.

Efficiency

Efficiency was confirmed as a crucial usability factor as we found that poor efficiency makes authentication concepts unfeasible for the mobile context. In this regard, various lab and field studies indicated an acceptable upper bound of roughly four seconds for the whole authentication process (orientation and input time). Even more important, we found that user's perceived efficiency often differs from measured efficiency values. For example, we found that *Marbles* was perceived faster than *Four Connect* even though it was measurably slower. A detailed analysis of the different stages of the authentication process indicated that the perceived efficiency is mainly determined by the relation between preparation effort and input effort. In particular, we found that high orientation times are more annoying than high input times. This indicates that the design of short preparation tasks is most important when optimizing efficiency. Overall, the results confirmed the importance of assessing all steps of the authentication process quantitatively and qualitatively. In addition, the performed field studies confirmed that efficiency strongly depends on the context. Therefore, it is important to assess the real-world efficiency of a system outside of the lab.

Perception

User perception has been shown to be a main factor for the acceptance of a concept. The observed differences in quantitative performance measures and qualitative user ratings indicate that measured performance is not always a good predictor for perceived performance. In addition, we found that the perception of contextual factors (e.g., risk level) plays an important role for user acceptance. Besides satisfaction, likeability was confirmed as an important factor. In this regard, the results indicated that improved user experience (e.g., better visual design) can directly increase performance ratings and showed that even early prototypes should provide a nice look and feel. We conclude that the detailed analysis of satisfaction and likeability are not a bonus to quantitative assessment but an equally important measure.

Memorability

Even though memorability was not systematically investigated, recall tests and qualitative error analyses indicated that memorability is less a problem in the mobile context. Due to the high frequency of use, unlock secrets are learned quickly, are continuously refreshed and are therefore seldom forgotten. However, we found that memorability is not only influenced by the type of secret but also by the type of interaction. For example, the analysis of randomized authentication concepts revealed that the same secret becomes harder to remember when the complexity of interaction is increased. On the one hand, this confirms that visual memory and motor memory are indeed important factors for the memorization of gestures. On the other hand, we learned that concrete representations of input elements can improve memorability by supporting story-based memorization when motor memory effects are not supported. We conclude that memorability is not a big problem as long as secrets can be created with a limited number of memorable chunks. Nevertheless, there is always room for improvement and memorability cannot be neglected.

Password Exposure

Password exposure was analyzed in terms of smudge attacks and observation attacks. The results confirmed the vulnerability of currently used authentication mechanisms. However, we found that such risks are rarely perceived as critical. As a consequence, most users are not willing to invest additional effort to be protected from password exposure. This implies that feasible concepts must remain very efficient and security must be provided by design. The systematic analysis of observation attacks revealed that observation resistance is influenced by a variety of factors including efficiency, interaction style and gesture composition. In contrast, smudge resistance is mainly influenced by the arrangement of input elements. Overall, we conclude that observation resistance and smudge resistance are important requirements for secure mobile authentication and gesture-based interaction design has great potential to improve the security of current systems. However, practical security should only be improved within the ranges of acceptable usability. As a consequence, designers should prioritize usability and rather focus on slight improvements instead of developing secure solutions which are not applicable to the real world.

Password Space

The systematic analysis of user-selected gestures provided important insights into the practical password space. We found that the relation of guessing resistance and input effort is a critical factor. Concepts which allow users to choose from a wide range of input complexity levels (e.g., Android unlock patterns) usually lead to a predictable selection of such secrets which are particularly easy to enter. Concerning the assessment of gesture strength, similarity was found as an important measure. While traditional composition aspects like length are indeed important, we illustrated that such metrics cannot directly be transferred from alphanumeric passwords to patterns. For example, it was shown that secret length has limited explanatory power in combination with grid-based gestures. The analysis of various concepts indicated that randomization can be applied in several ways to improve guessing resistance. We conclude that authentication mechanisms need to be designed in a way that increases the diversity of selected secrets. In addition to the optimization of the strength and the diversity of used secrets, aspects like an adequate theoretical key space and support of encrypted storage should be prerequisites.

5.1.2 Ten Concrete Recommendations for Future Designs

Each presented project provided valuable insights into the risks and potentials of gesture-based authentication on mobile devices. This Section summarizes the most important design objectives and gives concrete recommendations for the design of feasible authentication mechanisms. While some of the recommendations are consistent with general requirements of user interfaces [228], other recommendations are more specific to mobile authentication.

1. Maximize Security without Sacrificing Usability

Guessing resistance, smudge resistance and observation resistance are all valid and important security goals for knowledge-based authentication mechanisms. However, Chapter 2 and the analyses of this thesis illustrate that increased security often comes with reduced usability. Even though we learned that security can be improved in many ways, we claim that improved security is not a selling point and usability must not be sacrificed for security. In this regard, it is important to keep in mind that the mobile context makes greater demands on usability than other environments. First of all, the high frequency of authentication renders efficiency very important. In addition, since mobile device interaction is often a secondary task, the authentication should be exceptionally easy to perform. As a consequence, designers of mobile authentication methods must prioritize fast input, very short orientation times and quick error recovery. We conclude that concepts which are more secure but not usable (enough) cannot increase real-world security as there is no chance for wide user acceptance in the field even though such concepts might be rated acceptable in the lab.

2. Tailor the Interaction to the Mobile Context

As already mentioned, mobile authentication does not take place in the vacuum. Rather the context is a very important influencing factor for the design of mobile authentication mechanisms. Mobile devices are often used while the user focuses on a different main task (e.g., walking). Therefore, designers are required to reduce the visual, mental and physical load to a minimum. The results of this thesis showed that gestures are well suited to achieve these goals. *XSide* illustrated how gestures can help to minimize the load on the visual channel. As the system was designed for eyes-free interaction, users could even authenticate without looking at the device at all. Furthermore, we found that feasible solutions need to support one-handed interaction and authentication should be optimized for portrait orientation. For example, the evaluation of the *Pattern Rotation* schemes showed that users generally dislike physical rotation tasks. Finally, as users might not focus on the authentication, the mental effort must be kept very low. Gestures can support this goal by exploiting motor memory effects which reduce the cognitive requirements. We claim that all these factors are specifically important in the mobile context and might have different value in other scenarios (e.g., ATMs). It is important to note that this aspect renders field evaluations very important since the designed concepts need to be tested in the actual mobile context. We claim that lab studies have often limited explanatory power since authentication is usually presented as a main task and important influencing factors are not considered (e.g., physical movement).

3. Exploit the Advantages of Touch-based Mobile Devices

As stated above, the mobile context makes great demands on the usability of authentication systems. At the same time, mobile devices have specific characteristics which should be considered and exploited to improve security and user experience. The concepts proposed in this thesis were based on the assumption that mobile devices are (1) touch-based and (2) hand-held. The results showed that gestures are well suited for such devices. In addition, interactive touch screens enable different interaction styles and simplify the use of randomized spatial layouts or dynamic output. Such features enable developers to design authentication concepts which dynamically guide user input and adapt to the current situation. For example, *SwiPIN* illustrated that traditional concepts like PIN-entry can significantly benefit from touch-based interaction and dynamic guidance. In addition, hand-held devices enable the use of physical features like tilt and rotation. Designers should consider that mobile devices can be grasped in different ways and interaction can take place on different sides and positions. Especially *XSide* illustrated that such features can be exploited to increase security and at the same time maintain usability. Finally, even if it was out of scope of this thesis, mobile devices provide various different sensors like GPS, gyroscopes and cameras. Such sensors can be utilized to develop novel authentication concepts which utilize such features to analyze the user context or to support novel types of interaction.

4. Provoke Desired User Behavior by Design

Designers should not assume that users will use the system in the desired way. The results confirmed that gesture selection is often very predictable. In addition, we found that security is often jeopardized by users' input characteristics. Even though participants of the *SwiPIN* study were informed about the importance of not performing gestures directly on the shown digits, most users did it subconsciously. This behavior made *SwiPIN* vulnerable to observation attacks. However, we also demonstrated that unwanted user behavior can be effectively prevented by adjusting specific design factors like visual representations and interaction styles. We conclude that user behavior needs to be assessed in every stage of the development process to identify drawbacks in design. In this regard, designers should not assume that users consciously act in the most secure way. In contrast, feasible systems should provoke secure user behavior by design. This implies that the most usable way of interaction should accord with the most secure way of interaction.

5. Consider Providing Security on Demand

Increasing the security level of authentication methods often increases the complexity of input tasks. Even though all concepts presented in this thesis were specifically designed to provide good usability, a drop in performance was unavoidable. This is a critical issue as we found that mobile authentication often takes place in trusted environments where advanced security features like observation resistance might not even be necessary. As users are usually not willing to invest unnecessary additional effort, they often abstain from using security-optimized concepts and stick to insecure but more usable methods. We argue that adaptive security concepts represent a trade-off between security-optimized and usability-optimized solutions. By designing flexible authentication methods which support different security and usability levels, we can empower users to adapt the interaction to their current needs. Section 4.3 showed that adaptive security measures can be implemented as an add-on to established methods (e.g., *SwiPIN*) or as standalone authentication systems (e.g., *XSide*). We argue that such concepts increase user acceptance for two reasons: Firstly, authentication is very efficient in most situations where additional security is not required. Secondly, we assume that users are more willing to accept extra costs for more secure interaction when risks are actually perceived. However, it has to be noted that user-regulated security implies a higher vulnerability to attacks as users may not be able or willing to correctly adapt to a given risk level. We conclude that permanent security with high usability must be the main goal but adaptive security concepts can be a practical compromise as secure but inefficient methods are unlikely to be used in the wild.

6. Minimize the Number of Context-Changes and Interruptions

We repeatedly observed differences between measured and perceived performance. In-depth analysis of the tested concepts indicated that user satisfaction was downgraded by context-changes and interruptions. Concepts were perceived as less usable when the authentication process included different subtasks of varying difficulty. As a concrete example, we take a look at the results of *Pattern Rotation* and *Marbles*: While the *Pattern Rotation* scheme performed measurably faster than *Marbles*, it was perceived significantly slower. An analysis of the respective subtasks of the authentication concepts reveals one potential explanation for the phenomenon. While the authentication process of *Marbles* comprises identical subtasks (i.e., selecting a marble), *Pattern Rotation* is based on two different tasks. First, the user has to recognize the direction of the matrix, then the input is performed. Depending on the user's strategy, the first task (orientation phase) either increases the mental load (i.e., mentally rotating the matrix) or it increases the physical load (i.e., physically rotating the device). However, both strategies differ significantly from the subsequent input task. Similar effects have already been observed in neurobiological experiments [34, 293]. We additionally showed that the negative effect is stronger when the orientation effort exceeds the input effort (as in *Pattern Rotation*). At the same time, the results indicated that gestures are generally well suited to reduce the number of interruptions. Firstly, the input is often based on continuous movements which are perceived as efficient. Secondly, gestures allow informative but non-interruptive feedback by visualizing the user input. Finally, gesture-based input usually supports implicit abort and confirmation which reduces the number of subtasks and allows efficient recovery from errors. We conclude that besides minimizing the number of subtasks and interruptions, context-changes between subtasks should be avoided, meaning that mental or physical effort should remain at a constant level.

7. Minimize the Range of Input Complexity and Memorability

The analysis of user-selected Android patterns revealed that most users opt for simple shapes which are particularly easy to enter and presumably easier to recall than other gestures. Unfortunately, this makes the selection behavior very predictable and reduces the overall security of the system. User feedback revealed that gesture selection is mainly influenced by input complexity and memorability plays a secondary role in the mobile context. This indicates that whenever authentication systems provide a wide range of differently complex secrets, users are likely to opt for the lower range of complexity. Therefore, we argue that the range of complexity directly influences the size of the practical password space. *Android gestures* are a negative example, as a large portion of the theoretical gesture space is too complex and not suitable for daily use. We argue that mobile authentication mechanisms should be designed in a way that all available secrets are comparable in terms of input complexity and memorability. The *Marbles* concept illustrated that randomized spatial layouts can effectively decouple input complexity and security. As the order of input elements was randomized, input complexity was comparable for all possible secrets of the same length. The analysis of *Marble Gap* indicated that the interplay of input effort and gesture selection could also be utilized to systematically influence user choice. We conclude that feasible concepts should generally provide low input complexity and high memorability. In addition, providing a small range of complexity can be a way to increase the practical password space. If an authentication system supports a wider range of complexity, an adequate number of secrets should be available in the low-complexity sector.

8. Improve User Experience and User Acceptance by Aesthetic Design

It is common knowledge that people like aesthetic designs and prefer things which are visually appealing. We found that this is also true for security measures like lock screens. For example, we found that user acceptance increased for *SwiPIN* after the graphical user interface was redesigned. Although *Connect Four* and *Pattern Rotation* are based on similar principles, *Connect Four* was rated more satisfying. We argue that, similarly to other applications, users prefer authentication systems which are better designed. Since likeability is a strong influencing factor for satisfaction, even early prototypes should provide a good look and feel. We conclude that visual design is an important aspect at every development stage. By improving the user experience of authentication mechanisms, we can directly increase user acceptance. Similar findings have already been reported in connection with biometric concepts [70]. We claim that authentication systems should be designed as attractive as possible without hindering performance and that visual design should not be limited to the functional requirements.

9. Consider Supporting Established Secrets and Familiar Concepts

The evaluation of the first *Marbles* scheme revealed that people are not used to memorize secrets based on color combinations. In the second iteration, we changed the representation of the input elements and used concrete symbols instead of abstract colors. The results confirmed that this modification improved the usability as it enabled users to remember their secrets based on objects and stories. In addition, *SwiPIN* illustrated that established concepts like PIN-entry can be improved by slight modifications of the user interface. *SwiPIN* users can keep their familiar secrets which lowers the barrier to accept the new system. We argue that usability and accessibility is generally improved by supporting established concepts which users are already familiar with. However, depending on the design objective, it may not always be possible to utilize familiar concepts. For example, the preliminary evaluation of *XSide* showed that none of the established authentication concepts was feasible to enable back of device interaction. As a consequence, secrets had to be based on a novel type of gestures. We conclude that designers should try to minimize the differences to established concepts and evaluate if the design goals can be achieved by modifying established authentication methods before developing novel ways of authentication.

10. Reduce the Number of Required Authentications

The results of our field evaluation indicated that sensitive data is seldom accessed and that authentication often takes place in trusted environments. Even though the design of context-based authentication mechanisms was out of scope of this thesis, we argue that the number of required authentications should be reduced. First of all, designers of mobile authentication mechanism should reconsider the all-or-nothing access model. We assume that app-dependent and context-dependent models can help to reduce the authentication overhead. In addition, behavioral authentication mechanisms can help to strengthen the user-device binding and obtain an adequate security level without explicit authentication. We assume that such novel concepts can increase the security and the usability of mobile authentication. Nevertheless, there is no straightforward solution to the problem of unnecessary authentications as the context of the device is hard to define and data sensitivity is hard to measure.

5.2 Design Assistance for Gesture-based Authentication

This Section dissects the design space and illustrates the impact of specific design decisions. At the end of this Section, we will illustrate how the provided mapping can serve as a valuable resource for a goal-oriented design approach and thus help in further research projects.

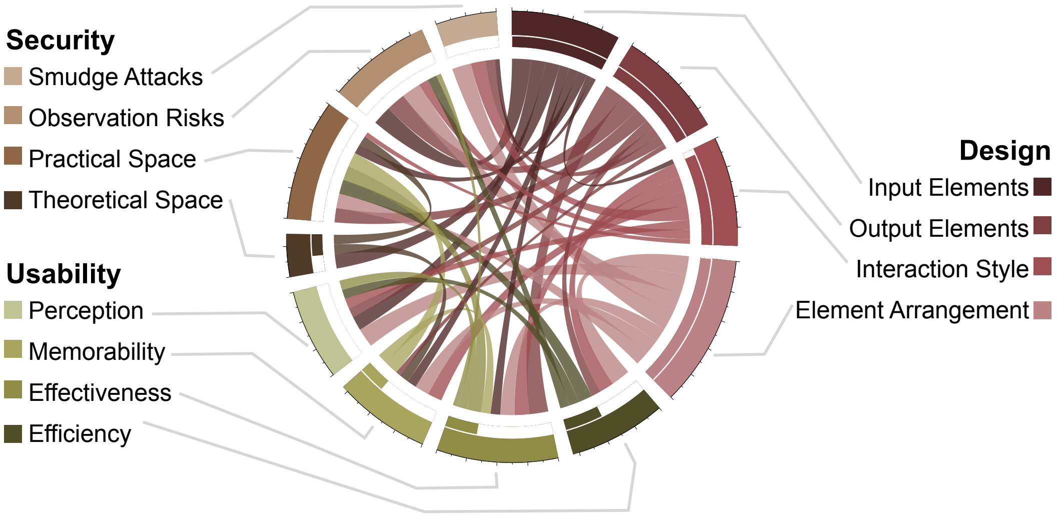


Figure 5.1: The circle chart maps the observed interconnections between design and problem space. Influences are represented by incoming and outgoing edges. The relative impact is represented by the width of the edges and the relative size of the respective sectors.

Figure 5.1 maps all influences observed and reported in this thesis. The visualization is based on a circular plot which was originally proposed to visualize migration flows [210]. However, similar visualizations have already been used in connection with graphical passwords [214]. The circle's segments are clustered using location and color and represent the main factors of the design and problem space. Observed influences are illustrated by incoming and outgoing edges. In addition, the relative impact of a factor was encoded using four width-levels ranging from minor to strong. The direction of a specific influence is encoded by the color of the influencing factor. Finally, the relative importance (i.e., many interconnections) of a factor can be derived from the size of the segment.

The visualization is cluttered and indicates a large quantity of interconnections. This Section will outline each segment (factor) in detail. Please note that we make no claim to completeness. We assume that besides these observed influences, there are various other interconnections between design and problem space which have not been evaluated. We will first discuss the impact of individual design factors before we illustrate how the mapping can help in designing new authentication mechanisms which meet the requirements outlined in Section 5.1.

5.2.1 Mapping Design Factors and Design Objectives

This Section revisits the main design factors and discusses their impact on usability and security. In addition, the respective impact will be visually highlighted using the circular chart presented above.

Input Elements

Input elements represent active areas which can be used to enter data. We specifically focused on the representation and the reusability of such elements. Figure 5.2 indicates that the specification of these aspects has strong effects on the problem space. In addition, we found interaction effects as some representation styles require specific interaction methods.

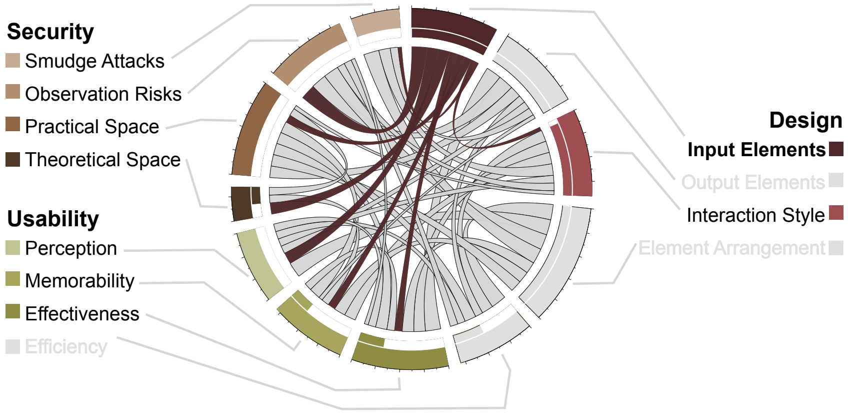


Figure 5.2: Input elements influence most considered aspects of the problem space. In addition, the representation of input elements can have influences on the interaction style.

Representation is categorized as “none”, “abstract” and “concrete”. We found that concrete representations of input elements are usually easier to recall than abstract tokens. The analysis of *Marbles* indicated that pure color-codes should be avoided as users are not familiar with memorizing such secrets. In contrast, concrete symbol-like representations were easy to remember. In addition, the representation style directly influences the user experience and nice visual designs are likely to improve satisfaction and user acceptance. Finally, input elements should be designed in a way that makes them easy to recognize and easy to distinguish to increase effectiveness.

The representation of input elements directly influences practical security. Firstly, it may influence the size of the practical password space as users may prefer specific representations over others. For example, Section 4 indicated that positive emojis are more likely to be selected than negative ones. Secondly, the representation of input elements influences observation resistance. To increase observation resistance, it is beneficial to design input elements in a way that makes them hard to track from a distance. The evaluation of *XSide* illustrated that giving up visual representations can significantly increase observation resistance. However, this design decision interferes with interaction style as target-oriented input becomes very hard without visual input elements.

We conclude that the main trade-off is found between memorability, user experience and observability. Abstract representations may lead to an acceptable compromise between the concrete representations (very usable) and no representation (hard to observe).

Reusability describes how often input elements can be activated (used). It is specified as "limited" and "unlimited". Limiting the reusability of input elements can increase effectiveness. For example, Section 3.3 revealed that enabling unlimited reusability of Android grid cells introduced novel types of errors and lowered the success rate. Therefore, limiting the reusability may be beneficial whenever input elements are positioned close to each other and users are likely to activate input elements accidentally. On the other hand, unlimited use of the same input element may increase input speed as using the same element multiple times enables faster input than using different elements. This is especially true when distances between different input elements are larger. However, since this factor was not assessed in our studies, it is not included in Figure 5.2.

Furthermore, reusability represents an important security factor. First of all, the reusability of input elements influences the upper bound for the theoretical password space. If the reusability of input elements is limited, designers need to provide a larger number of input elements to achieve high theoretical security. In addition, the evaluation of *Marbles Gap* indicated that reusability can be utilized to influence the size of the practical password space. Limiting the number of inputs for a single element can nudge users to use different elements and can diversify password selection. Finally, Section 4.2 indicated that the reusability of input elements affects the vulnerability to smudge attacks. The order and type of input elements is easier to deduce when input is limited to one (e.g., Android unlock gestures).

We conclude that, related to reusability, there is an inherent trade-off between effectiveness and password space. However, no clear recommendations can be given as the effects of reusability strongly depend on the specification of other design factors.

Output Elements

Output elements represent all visual elements which are not interactive. We distinguished between guidance and feedback elements. Figure 5.3 indicates strong influences on password space, observation resistance and effectiveness.

Guidance is optional. If guiding elements are present, they are distinguished into static and dynamic approaches. Similar to visual input elements the visual appearance of guiding elements has strong influences on user perception. Especially static guidance is often used to support the recall of secrets. Such cues may give hints which help users to remember their gestures. Efficiency and effectiveness is directly affected whenever guidance is required to

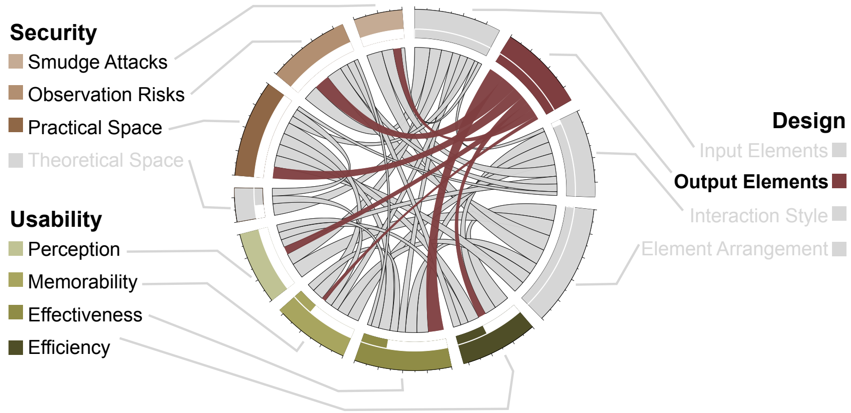


Figure 5.3: While output elements have a strong impact on the problem space, other design factors are rather independent.

successfully authenticate. We presented several approaches which strongly relied on guiding elements. *Pattern Rotation* used static guidance to indicate the direction of the grid, *SwiPIN* used dynamic guidance to communicate the gesture mapping. In such cases, where user interaction depends on guidance, the output elements must be designed in a way that makes the guiding cues quick and easy to understand. Finally, dynamic guidance tends to slow down the authentication process whenever users have to react to such cues in real time.

Nevertheless, we illustrated that suitably designed guidance allows for effective use of randomized authentication mechanisms. Therefore, guidance is especially useful to develop systems which provide increased protection from smudge attacks and observation attacks. In addition, we showed that selection behavior can be effectively influenced by guiding elements. For example, Section 4.4 indicated that guiding output elements can nudge users towards the selection of specific grid cells.

We conclude that suitably designed user interfaces are usually not in need of guiding elements. However, guidance can be utilized in various ways to increase practical security.

Feedback is optional. If feedback is provided it can be based on aggregated or detailed information. It is common knowledge of user interface design that comprehensible feedback increases efficiency and effectiveness [228]. However, the analysis of current mobile authentication methods revealed specific aspects of feedback on errors. We found that failed gesture input is usually communicated via detailed feedback. This feedback is often given by visualizing the entered path of a gesture. As a consequence, users can easily understand which aspect of the gesture was wrong. In addition, errors are detected earlier when feedback is provided. Therefore, visualizing the entered path makes error recovery very efficient, an aspect which is very important in the mobile context. In addition, we assume that providing detailed feedback during input can increase memorability as it supports a visual coding of the secret. Finally, feedback communicates interactivity and can improve user experience.

On the other hand, we found that feedback is an important design factor in terms of security. First of all, it influences the observability of the system. Section 3.4 revealed that grid-based gestures are significantly easier to observe when gesture input was visualized. In addition, we found that detailed feedback on errors can increase observation risks. For example, Android unlock patterns give detailed feedback on input errors and visualize the entered gesture even if input visualization has been switched off. This can expose parts of the user's gesture. Finally, Section 2 illustrated that feedback can be applied to improve gesture selection (e.g., strength meters).

We conclude that detailed feedback is preferable in terms of usability but it may increase observation vulnerability. Therefore, designers should consider to provide aggregated feedback if possible. Finally, visual feedback may be omitted when it can be substituted by functional feedback. For example, unlock screens do not need feedback on successful unlock events as the feedback is given implicitly by unlocking the device.

Interaction Style

Gesture-based interaction was categorized according to the relation of the gesture and the input element and according to the directness of the interaction. Figure 5.4 indicates strong influences on usability and practical security.

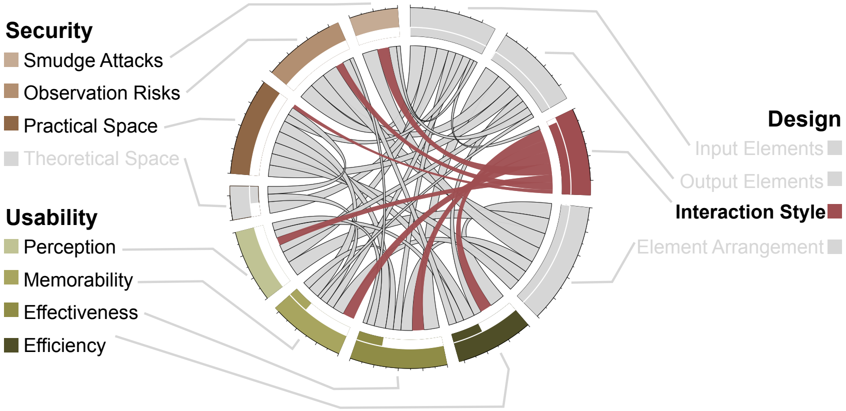


Figure 5.4: The interaction style characterizes the type of authentication system. It influences all factors which are directly associated with the input process.

Relation specifies if a gesture is self-contained or target-oriented. First of all, the relation of a gesture influences how gestures are retrieved from memory. Self-contained gestures often exploit motor memory effects, while the memorization of target-oriented gestures is often focused on the targets themselves. *XSide* represents an example for a self-contained gesture concept which is mainly memorized in form of movements. On the other hand, *Marbles* is a fully target-oriented concept which uses gestures solely for input, meaning that the gesture itself is not part of the secret. However, as illustrated by Android unlock patterns, target-oriented concepts can also be designed in a way to support motor memory effects. Naturally, the interaction style influences effectiveness and efficiency. For example, *XSide* illustrated that self-contained gestures enabled effective and efficient eyes-free interaction. *SwiPIN* showed that target-orientation can increase usability as the use of two input elements allowed the use of a reduced and therefore simplified gesture set.

Furthermore, gesture relation can have strong effects on practical security if it is suitably combined with other design factors. For example, *SwiPIN* illustrated that target-oriented gestures can improve observation resistance as users can be forced to perform the gestures at specific regions of the screen. On the other hand, self-contained gestures which allow eyes-free interaction can be used to authenticate on the back-of-device and therefore increase observation resistance. Finally, the relation of gestures can directly influence smudge resistance and gesture selection depending on the design of the interaction concept itself.

We conclude that target-oriented and self-contained gestures can be used in various ways. However, target-orientation is needed if users shall be nudged to specific regions or specific input behavior. Even though target-oriented gestures are the standard approach for established concepts, self-contained gestures should be considered as they provide desirable features in the mobile context.

Directness describes if gestures are performed directly on the input element (target) or indirectly on another position. Directness is very likely to influence user perception. The evaluation of *SwiPIN* indicated that indirect input should be designed in a way that clearly separates input and output areas. If indirect input takes place near the actual target, users may feel forced to use a specific input position. On the other hand, we observed that indirect gesture input can positively influence effectiveness and efficiency as separating input areas and target areas reduces occlusion and fat finger problems.

In terms of security, we found that the directness of interaction influences smudge attack vulnerability and observability. Direct input makes smudge attacks specifically effective if it is combined with fixed spatial arrangements. In such cases, the smudge traces on the touchscreen tend to match the displayed input elements. In contrast, indirect interaction can decouple activated target elements and the traces of interaction. In addition, Section 4.3 illustrated that observation resistance can be increased by indirect interaction as bystanders need to observe two distinct areas simultaneously.

We conclude that directness is an important factor to increase the practical security level of an authentication system. While direct interaction is the obvious choice for what probably feels most natural for the user, the results showed that indirect interaction can also improve usability factors and should be considered.

Element Arrangement

Element arrangement describes how input and output elements are arranged on scree. It is distinguished into spatial and temporal arrangement. Figure 5.5 indicates that element arrangement relatively is the strongest of all factors (i.e., the largest segment).

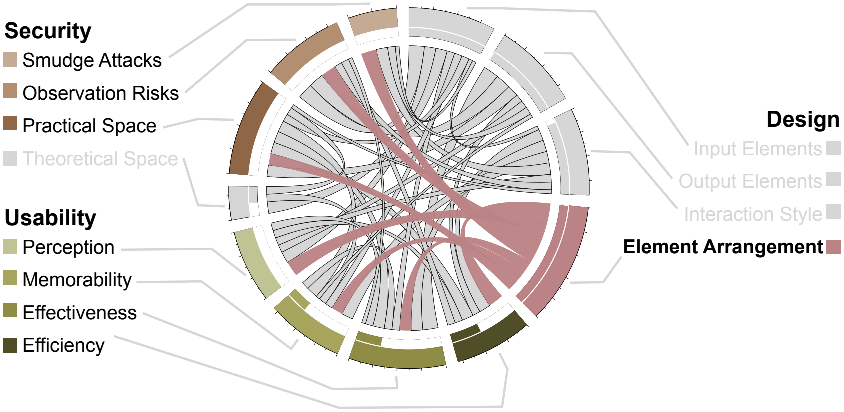


Figure 5.5: Element arrangement has a significant impact on usability and security. It is therefore the main factor to adjust the trade-off between both design goals.

Spatial arrangement describes how elements are positioned on the screen. We distinguish between fixed and random arrangements. Randomization tends to increase error rates and authentication times while fixed layouts support motor memory effects and short orientation times. Therefore, fixed layouts are preferable in terms of usability. If randomization is applied, it should be designed in a way that allows efficient orientation and fast input. In addition, clear guidance should be provided to allow for effective orientation and short search times. Finally, randomization has a strong impact on user perception. We found that randomized interfaces are particularly annoying when the orientation task makes high demands in terms of mental effort or physical effort. Especially the evaluation of *Marbles* and *Pattern Rotation* indicated that randomized spatial arrangements benefit from splitting the orientation task into multiple sequential challenges as the randomization was well accepted in such cases.

While randomization should be avoided in terms of usability, it can increase security in various ways. First of all, Section 4.2 showed that a randomized spatial arrangement of input elements can effectively fend off smudge attacks. Secondly, *SwiPIN* illustrated that randomized position of output elements can be used to protect PIN-entry from observation attacks. As the output elements were randomized but the input elements were fixed, the system did not require extensive visual search tasks and enabled very short orientation times. However, we also showed that observation resistance can be increased without randomizing input or output elements. *XSide* provides two fixed input elements and therefore supports fast authentication and good memorization. Finally, spatial randomization can have a strong influence on the practical password space. On the one hand, fully randomized concepts can partly decouple input effort and secret composition. As the positions of elements constantly change, users have no benefit from prioritizing specific regions on the screen and may select a more diverse set of elements. Secondly, Section 4.4 illustrated that randomized positioning of presentation effects can effectively influence gesture selection even if authentication is based on a fixed layout.

We conclude that fixed spatial layouts should be preferred whenever possible. When randomization is applied it should be designed in a way that minimizes orientation effort. Mixed approaches with fixed input elements and randomized output represent a trade-off in terms of usability and security.

Temporal arrangement describes the number and order of distinct input tasks. According to Schaub et al. [214], we distinguish between single challenges and multiple challenges. Single challenges are usually more efficient as users can authenticate with one continuous gesture. On the other hand, multiple challenges can be useful to split longer and more complex tasks in simpler and shorter subtasks. The evaluation of *Marbles* showed that authentication concepts which are based on multiple very easy subtasks can be very effective and easy to use. In addition, we found that the temporal arrangement has a very strong influence on user perception. Section 4.2 indicated that context changes between multiple challenges should be avoided and difficulty should remain constant to increase user satisfaction. This is especially important if multiple temporal arrangements are combined with randomized spatial arrangements.

In terms of security, we observed strong effects on observation resistance and smudge attack vulnerability. This was especially the case, when randomized spatial layouts were used. *SwiPIN* exploits the fact that subtasks (i.e., entering one digit) are too short to allow observers to make sense of the presented mapping. *XSide* supports a very flexible input style as some challenges can be performed on the front side and others can be performed on the back of the device. This additionally increases observation resistance as attackers are likely to observe only a subset of the authentication task.

We conclude that quick and easy single challenges are generally preferable. However, usability can benefit by splitting complex authentication task in multiple easier subtasks. In addition, multiple challenges can help to improve security in terms of smudge-resistance and observation-resistance. When multiple challenges are required, it is important to avoid context changes between these challenges as such changes negatively influence user perception.

Further Influencing Factors

In addition to the impact of the design factors, we observed various interaction effects between factors of the problem space. Knowing of such interaction effects is crucial to be able to set the right design goals. Figure 5.6 indicates that usability has strong effects on security.

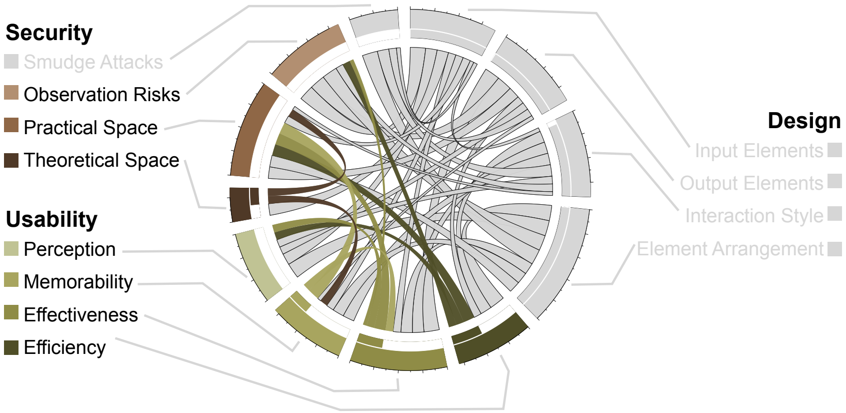


Figure 5.6: In addition to the actual design factors, efficiency, effectiveness and memorability have strong influences on the practical security of authentication systems.

Usability factors have a strong influence on security. This again confirms that there is no clear trade-off between usability and security but usability is a precondition for security. As already discussed in Section 2, memorability influences the practical password space as users prefer secrets which are easy to remember. However, in the mobile context efficiency and effectiveness are very important factors, too. The analysis of grid-based gestures indicated that users select secrets which are particularly fast and easy to enter. In addition, we found that effectiveness and efficiency can influence observation resistance. Firstly, fast user interaction is harder to observe. Secondly, input errors increase the number of possible observations which increases the success rate of observation attacks. Finally, user perception is strongly influenced by the efficiency and effectiveness of a system.

Security factors rarely showed interaction effects. Nevertheless, the type and structure of theoretically available passwords naturally influences the practical password space and the memorability of secrets.

We conclude that high performance should be prioritized in the mobile context. The theoretical password space should provide a large number of easy to remember and easy to enter secrets to achieve high practical security.

5.2.2 Recommendations for a Goal-oriented Design Process

The previous Section mapped various interconnections between design and problem space. We claim that the presented insights can serve as a tool which enables a systematic and goal-oriented design process for novel authentication systems far beyond the specific constraints of this thesis. While the map was created focusing on mobile devices, we argue that the insights can be useful in various other contexts. The only precondition to use the tool is a clear specification of the design goals. As we illustrated, the mobile context makes great demands on efficiency. However, other authentication scenarios may have different requirements. Therefore, it may be necessary to analyze the problem space of the respective use case to identify proper design objectives. As illustrated in Figure 5.7, we outline two potential application areas for our map: Bottom-up design and top-down analysis. The bottom-up approach starts with the definition of individual design factors while the top-down approach starts with analyzing an existing concept. The following text gives an overview of both approaches.

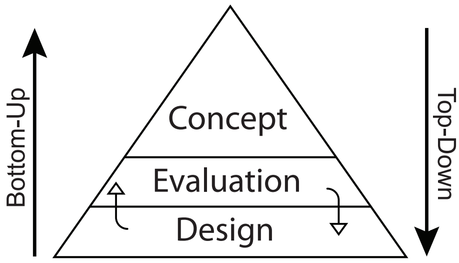


Figure 5.7: The presented insights are useful for both bottom-up design and top-down analysis.

Bottom-Up Design

The bottom-up design approach is especially useful to develop novel authentication systems from scratch. The development process often starts with a creative exploration of different design alternatives. As soon as the design objectives are specified, designers can systematically explore the design space and identify the most promising characteristics for each design factor. The design candidates are then improved based on an iterative process including several design and evaluation phases. We argue that a systematic design process is useful to tailor concepts to specific requirements and helps to avoid design errors a priori.

Example: To illustrate the approach, we give a simplified example. Let us assume that we plan to design a novel authentication mechanism for a specific application area. The analysis of the respective problem space has revealed that observation attacks do not occur in the considered environment. However, we found that a feasible concept must be very usable and should support eyes-free interaction. Based on these aspects, we specify the following desired characteristics of the design space:

Input Elements As observability is not a problem, we specify that memorability shall be optimized by using concrete representations of input elements. To reduce input errors, we furthermore decide that reusability of such input elements shall be limited to one.

Output Elements As dynamic guidance often requires visual focus, we decide to use static guidance instead. In addition, detailed feedback shall be given to support efficient error recovery.

Interaction Style As the authentication system shall support eyes-free interaction, we opt for self-contained gestures which shall be performed directly.

Element Arrangement To improve usability and to enable eyes-free interaction, we provide a fixed spatial arrangement of elements. In addition, authentication shall be performed within a single challenge to further improve efficiency.

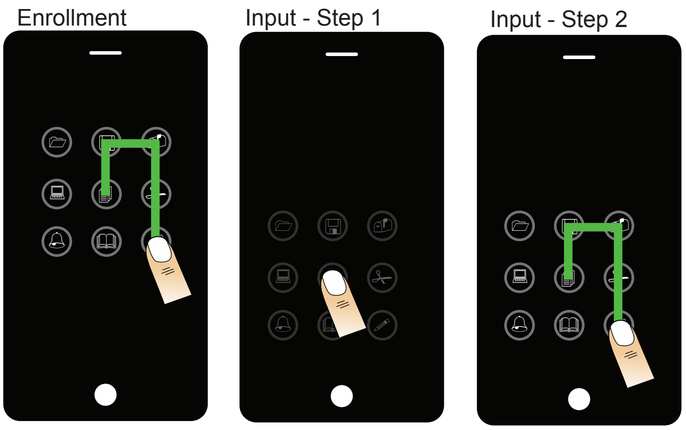


Figure 5.8: An exemplary illustration of a possible concept. The visual design is similar to established grid-based approaches. However, the approach supports eyes-free interaction as input is based on self-contained gestures.

One potential solution is illustrated in Figure 5.8. The concept is based on a visual grid which provides nine concrete elements. Interaction is similar to established grid-based concepts but less target-oriented. In contrast, the concept supports self-contained gestures which can be performed anywhere on the screen. As a consequence, the visual grid serves as a static guiding element and not as an input element. The position of the grid is derived from the location of the first touch event. As soon as the user touches the screen, the grid is presented with the starting cell positioned under the user's finger (see 5.8, center). The starting position of the gesture is stored during enrollment. Eyes-free interaction is facilitated by the fact that user input is not required to match the position of the grid as gestures are analyzed based on relative direction changes. However, by visually representing the grid, the system supports dual coded learning which might improve memorability. In addition, the system can be used in the same way as established authentication systems which may increase user acceptance.

This is just one possible design and it certainly has limitations. For example, authentication may fail if the user starts too close to the border of the screen. In addition, the presented concept is likely to negatively influence gesture selection. We assume that users will select simple shapes. However, the example illustrates how the gained insights can be used to systematically develop concepts from scratch. While such a systematic approach may simplify the ideation, it is important that candidate concepts are thoroughly evaluated in the next iterations. The results of user studies will help to reconsider design decisions and are required to further improve the concept.

Top-Down Analysis

The top-down approach is useful to identify possible areas of improvement in a given authentication concept. Therefore, the approach starts with analyzing a working authentication mechanism. Field evaluations of existing authentication systems often indicate potential problems. For example, we showed that established concepts are easy to observe. We argue that in such cases, it is beneficial to dissect the respective authentication system and compare the current state of individual design factors to their goal state. The goal state of the respective design factors can be derived from previously defined design objectives. On the other hand, top-down analysis can serve as an analytic tool which can supplement time-consuming and cost-intensive field studies. For example, the analysis of the individual design factors can point out potential drawbacks before the user study takes place.

Example: The approach is again illustrated by giving a simplified example. Let us assume that we plan to improve the standard PIN-entry. The evaluation of the problem space has indicated that PINs are well suited for the respective use case. However, we found that eyes-free interaction would be beneficial as it could improve usability and even security[1]. We would start by analyzing the standard PIN approach and specify its design factors:

Input Elements PIN concepts use abstract input elements which allow an unlimited number of inputs.

Output Elements Static guidance is usually given by illustrating the edges of the active input areas. In addition, most PIN concepts provide feedback when buttons are pressed and communicate errors in form of aggregated feedback.

Interaction Style The input is based on direct target-oriented interaction.

Element Arrangement PIN concepts present input elements in a fixed spatial arrangement. Authentication takes place in multiple challenges, whereby each subtask is described by entering one digit.

Figure 5.9 illustrates one possible solution. The analysis revealed that target-orientation is the main limitation in terms of eyes-free interaction: To enter a four-digit PIN, users need to hit four small targets. Therefore, we might come up with the idea to provide larger input areas. To provide a familiar look-and-feel, the additional input areas have no representation. Figure 5.9 (left) illustrates one possible layout: Outer areas are positioned directly at the borders of the device and center areas are enlarged. We assume that both aspects facilitate eyes-free interaction. On the one hand, hitting the larger inner buttons does not require precise pointing. On the other hand, the border of the device serves as haptic guidance and facilitates hitting the outer buttons. Figure 5.9 (right) illustrates a user entering a “0”.

[1] Observation resistance could be increased as authentication could easily take place out of sight.

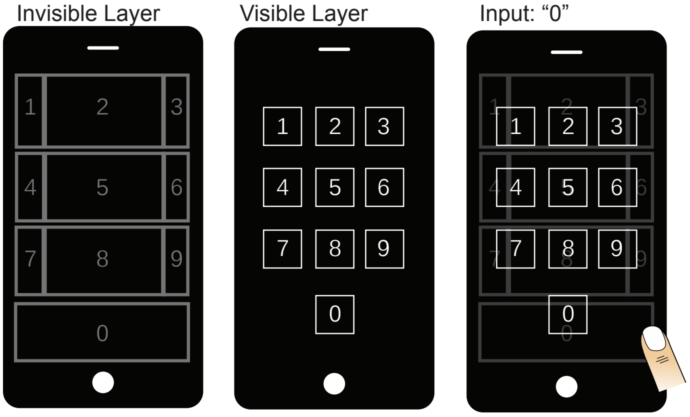


Figure 5.9: An exemplary illustration of a possible concept. While the appearance matches the PIN concept, eyes-free interaction was facilitated by adding a second layer of input elements.

Based on a top-down analysis, we would assume that the alternative interaction concept is more secure against observation attacks. In addition, we would assume that it is perceived slower than the original PIN-entry as input complexity differs between some digits (subtasks). Finally, we would assume that users prefer the standard PIN pad whenever they are able to look at the screen. As a consequence, we would recommend to support both concepts. To avoid unwanted interference between input modes, we could provide an explicit switching mechanism. Nevertheless, it is important to note that systematic top-down analysis cannot substitute user studies and empirical evaluation.

5.3 The Evaluation of Mobile Authentication Methods

The research projects presented in this thesis provided valuable insights into the evaluation process of mobile authentication mechanisms. While the recommendations presented in this Section were gathered in connection with gesture-based mobile interaction, they are not limited to the evaluation of gesture-based approaches. In contrast, we claim that the insights are also useful for the evaluation of authentication mechanisms in other contexts (e.g., public displays). While Chapter 3 presented mainly descriptive and exploratory basic research, Chapter 4 presented more explanatory research projects. This Section presents the research methods according to their main purpose. Please note that in practice, research projects have mixed goals and explanatory studies often comprise exploratory elements. While we specifically focus on the evaluation of authentication mechanisms, Kjeldskov and Paay [153] provide a general overview of HCI research practices in the mobile context.

5.3.1 Exploratory Studies

Exploratory studies help to understand the research problem. Therefore, they are especially important at the beginning of a project to understand the authentication context and to define the research objectives. Over the course of this thesis, we used brainstorming sessions, focus group discussions, low-fidelity paper prototypes and pilot studies.

Brainstorming sessions are an important part of the ideation process as they help to come up with new ideas for both authentication concepts and evaluation strategies. However, we found that two aspects are critical for the outcome of such sessions: Firstly, it is beneficial to invite participants with superior knowledge of the discussed topic. Even experienced smartphone users are usually not able to correctly assess usability and security and thus ideas are rarely feasible. Secondly, the topic should be well defined. For example, instead of asking for "observation resistant" concepts, it is beneficial to present additional requirements like desired interaction methods, performance needs or context of use.

Focus group discussions are useful to evaluate already specified authentication concepts or evaluation strategies. In contrast to brainstorming sessions, we found that it is beneficial to invite end users to such discussions. If possible, concept ideas should already be presented in a visual way. Focus groups are a powerful tool to evaluate user perception and user behavior at an early stage. By discussing real world problems, researchers can identify important design factors and determine independent and dependent variables for later evaluations. Therefore, the most important topics at this stage include authentication behavior, risk perception and authentication context.

Paper prototypes are useful to identify design issues and to inform later study designs. We found that paper prototyping is very useful in connection with mobile authentication mechanisms. Due to the exploitative character of the concept development process, researchers often come up with various design alternatives. A thorough paper prototyping study can help to reject some of the concepts before they are implemented. In addition, concept design can be improved based on user feedback. It is beneficial to build interactive paper prototypes which allow for the simulation of enrollment tasks and authentication tasks. Data collection should focus on qualitative feedback. However, low-fidelity prototypes can also be useful to gather first quantitative results (e.g., password selection). Interaction should be filmed for later analysis. When testing multiple concepts, the order should be randomized.

Pilot studies are essential to identify critical issues in the study design or with the concept itself. When planning a lab study, the whole procedure should be tested at least once. When planning a longitudinal field study, concepts should be tested for several days. The preliminary user studies will help to identify issues before a time consuming and cost intensive study takes place. Special attention should also be given to automatic log files which need to be checked for completeness and correctness.

5.3.2 Descriptive Studies

Descriptive studies are essential to understand the current state of mobile authentication. While exploratory studies are mainly useful to understand basic usability aspects, descriptive analysis can effectively inform security aspects. Data collection should include both qualitative and quantitative data. Chapter 3 presented different types of descriptive evaluation. We found that ethnographic studies and controlled field studies are especially useful to investigate usability and user behavior while online studies are feasible to analyze basic security aspects.

Ethnographic studies are crucial to analyze user behavior and risk perception in the wild. Section 3.2 illustrated that the users' mobile devices can be utilized to enable long-term field observations. Describing and understanding real world aspects is important for the development of adequate solutions. While quantitative data (e.g., authentication frequency) can be automatically collected, qualitative factors are more difficult to assess. Experience sampling was shown as one adequate way to augment quantitative data with context information. We found that usability aspects can be directly assessed, while the description of security aspects is often based on user reported data which needs to be handled with care. For example, Section 3.2 revealed that observation risks are rarely reported. However, this does not directly indicate that observation risks do not exist. Another potential way to assess user behavior in the wild is based on field observation. For example, observation risks could be quantified by counting observation attacks in public spaces.

Experimental studies can be useful to describe the differences between different approaches (e.g., authentication concepts). While experimental lab studies usually have an explanatory purpose, we found that controlled field studies are feasible to describe the status quo of current authentication. In contrast to natural studies, controlled studies allow a more accurate assessment of quantitative data. However, increasing internal validity tends to decrease ecological validity. Therefore, the trade-off between control and natural behavior needs to be well considered. Section 3.3 presented a controlled field study which described the differences between grid-based gestures and PIN. While the study was performed in the natural environment, we had to control several real-world factors to allow for accurate measurement. For example, we controlled password choice and authentication frequency. We conclude that such field experiments provide a good trade-off between controlled lab studies and natural field studies. However, while the study type is well suited to evaluate usability aspects, we found that security remains hard to assess.

Online experiments are well suited for descriptive evaluations and enable large-scale studies with limited funds and within a limited amount of time. Since authentication tasks are usually based on short interactions which are relatively easy to describe, they are often easy to simulate in online experiments. We showed that online experiments are especially useful to describe security aspects. For example, Section 3.4 illustrated how gesture-input can be simulated to analyze observation risks. With the available resources, it would not have been possible to perform such an experiment in the real world. However, since online experiments limit both internal validity and ecological validity, real-world experiments should be preferred whenever possible.

Self-reported data is crucial to assess user perception. In addition, it is often the only way to assess password selection in the mobile context. As mobile secrets are stored locally on the user's device, databases of real-world passwords are usually not available. Section 3.5 presented a large-scale analysis of user-selected gestures which had been collected online. We found that simulating the enrollment process of current devices is a feasible approach to collect realistic secrets. For this purpose, the interface should mimic both the interaction and the visual appearance of the original system. Finally, when collecting such data, it is beneficial to ask participants for the truthfulness of their provided answers. For example, collected passwords can be excluded from analysis if participants indicate that their input was unrealistic.

5.3.3 Explanatory Studies

Explanatory studies are usually based on the results of exploratory and descriptive evaluations. Instead of describing a phenomenon, explanatory research aims at understanding the reasons of a phenomenon. Chapter 4 presented several types of explanatory studies. Notwithstanding the concrete study design, we found that explanatory studies benefit from following three aspects. Firstly, data should be collected on a high detail level. For example, user feedback should be collected using open-ended questions. If the analysis of open-ended questions is too costly, ordinal scales should be preferred to binary scales (yes, no). Secondly, usability and security needs to be precisely defined. Finally, it is important to design user studies in a way that both qualitative and quantitative data can be assessed. For example, counting authentication errors is as important as analyzing the source of errors.

Lab experiments are the most important research method to analyze the impact of different design decisions in a controlled environment. The prototype can be realized as a standalone application and optimized for a specific mobile device. This facilitates the development process and increases the level of reliability. The most important aspects to consider include the study design, the assignment of passwords and the kind of baseline condition. We found that a repeated measures design is well suited to analyze the usability of an authentication system and to analyze practical security aspects like observation resistance and smudge attacks. However, Section 2.5 indicated that between-groups designs are beneficial when password selection tasks are required and tasks are very similar. Furthermore, password assignment can have a strong impact on measured performance. We recommend to use different secrets of similar complexity at this stage. In later field analyses, self-selected passwords may be allowed. Finally, it is crucial to include the right baseline condition. We used the most popular gesture-based system and calibrated secrets in a way that the theoretical password space was comparable.

Overall, we found that lab experiments are well suited to collect performance data. However, researchers need to be aware of the fact that the observed performance may differ significantly from field performance. During lab evaluations, mobile authentication represents the primary task. In the wild, mobile authentication often represents a secondary task. In addition, novelty effects and Hawthorne effects are usually stronger during short-term lab tests than during long-term field evaluations. We assume that performance is likely to drop in the field. In contrast to usability, security is often well suited for lab experiments as such experiments often simulate worst-case scenarios. As long as the worst-case scenario can be evaluated in the lab, there is no need to perform field studies.

Field experiments are crucial to assess the real-world performance of an authentication system and to get realistic insights into user acceptance. To achieve the highest level of ecological validity, participants should use their own devices and the tested authentication concept should be implemented as real lock screen. As already mentioned, field experiments should focus on usability issues as security levels can be simulated in the lab. Participants should use the system for at least two weeks. To get a better understanding of the context, experience sampling methods should be applied. Finally, researchers should be aware of the fact that novel authentication mechanisms are shown around by participants and not every logged authentication event is valid. This makes the assessment of real-world behavior and real-world performance a major challenge.

Self-reported data is crucial for explanatory research. It is important to understand user perception, authentication context and usability issues. Especially the inductive coding of open-ended questionnaires can be very useful to get deeper insights into the effects of an authentication system. In addition, user feedback should be collected during briefing and debriefing. Finally, short in-situ questionnaires help to get instant feedback after authentication events.

Online experiments have limited value in explanatory research and should rather be used for descriptive studies. Section 2.5 showed that online experiments can be used to systematically compare different conditions. However, researchers need to collect a very large set of data to show significant effects. We argue that online experiments can serve as a last resort when other experiments are not feasible due to time constraints or money constraints.

Chapter 6

Conclusion and Future Work

Mobile use is growing faster than all of Google's internal predictions.

– Eric Schmidt, Executive Chairman of Alphabet Inc. (2011) –

This thesis investigated the risks and potentials of graphical gesture-based authentication mechanisms on mobile devices. We provided an in-depth overview of related work and presented results of diverse empirical research projects. The investigation of both established methods and novel solutions gave us a detailed understanding of the challenges and the opportunities of gesture-based authentication on mobile devices. This Chapter summarizes the main contributions and gives an outlook on future work.

Section 6.1 revisits the main research questions stated in Chapter 1 and summarizes the main contributions of this thesis. Section 6.2 provides four concrete perspectives for future work concerning authentication on mobile devices before Section 6.3 presents some final remarks.

6.1 Summary of Contributions

The thesis analyzed the risks and potentials of graphical gesture-based authentication mechanisms on mobile devices. We provided answers to the following main research questions:

RQ1 How do established gesture-based concepts perform in terms of usability and security [current state]?

RQ2 What are the requirements for improved authentication on mobile devices [goal state]?

RQ3 How must graphical gesture-based concepts be designed (and evaluated) to meet the requirements of mobile devices [process] ?

The analysis was based on fundamental field studies and applied design-oriented research and contributed to the understanding of mobile authentication at various levels. *RQ1* was mainly addressed by analyzing the usability and the security of established mechanisms in the field. We investigated the authentication context, the usability of PIN and gestures and the observability and guessability of gesture-based approaches. *RQ2* was approached by both field studies and the design of novel authentication concepts. Finally, we gathered enough insights to give concrete recommendations for the design and evaluation of feasible authentication mechanisms *(RQ3)*. This Section summarizes the main contributions of the thesis and illustrates their usefulness beyond the scope of gesture-based interaction and mobile devices.

Understanding the Requirements of Mobile Authentication

Chapter 2 illustrated that despite the diverse development of novel authentication mechanisms, the actual requirements for feasible authentication mechanisms had not been investigated. The lack of understanding of real-world factors often led to impractical solutions which were indeed more secure but not usable enough. This thesis contributed by specifying the problem space of mobile authentication mechanisms and by systematically investigating the relevance of individual factors.

Section 3.2 shed light on risk perception and unlocking behavior in the wild. The results showed that authentication methods compete with alternative protection strategies. Instead of relying on authentication, many users try to keep the device physically protected. We learned that acceptable authentication mechanisms must be very efficient and lacks of performance cannot be justified by increased security levels. Section 3.3 gave empirical evidence for the benefit of gesture-based interaction on mobile devices. In addition, we showed that the mobile context renders efficient error recovery more important than error prevention. The design of novel authentication mechanisms revealed further important requirements for usable solutions. For example, Section 4.2 illustrated the importance of short orientation times and showed that perceived usability is even more crucial than measured performance. In addition, Section 4.3 illustrated the importance of usability features like eyes-free interaction or one-handed interaction.

We further contributed to a better understanding of security factors and showed that practical security is an often unmet requirement. Section 3.4 quantified the observation resistance of established gesture-based passwords and showed that currently deployed concepts are very much prone to observation attacks. Section 4.2 showed that the same is true for smudge attacks. In addition, Section 3.5 quantified the practical password space of user-selected gestures and indicated that established methods are vulnerable to dictionary attacks.

Finally, Section 5.1 summarized the gathered insights and presented a revised version of the problem space. In addition, we derived concrete directions for future developments which support the specification of appropriate design goals and support the development of feasible solutions. While we assume that the presented requirements are specific to the mobile context, we argue that the general approach can be applied to any other authentication scenario. We argue that a thorough understanding of the authentication context is a prerequisite for any development of feasible authentication mechanism. We conclude that the presented problem space definition can be used as an expandable basis for the investigation of the specific requirements of other authentication scenarios.

Usable and Secure Authentication Methods

The thesis contributed by illustrating the design and evaluation of novel usable and secure authentication methods. Overall, we presented 17 concept ideas, out of which six concepts were fully developed and extensively tested. By reporting both good and bad design decisions, we contributed to an in-depth understanding of individual design factors.

Section 4.2 illustrated how gesture-based interaction can be designed in a way that smudge traces are hard to interpret. We presented general approaches to prevent smudge attacks and designed practical solutions. The two most promising solutions, namely *Marbles* and *Connect Four*, were finally implemented as lock-screen replacements and tested in the field. The results showed that both concepts were usable and significantly more secure. *Marbles* was actually less error-prone than the insecure baseline approach. In addition, we gathered valuable insights into the feasibility of randomized spatial arrangements and their impact on performance and perception. Section 4.3 illustrated the development of two observation-resistant authentication mechanisms which are fast and easy to use on mobile devices. We presented *XSide*, a gesture-based system which enables back-of-device authentication and *SwiPIN*, an observation-resistant input mechanism for PINs. Both systems were designed in a way that enables the user to adjust the usability-security trade-off to the current situation. Both concepts were significantly more secure but yet usable. *SwiPIN* was actually selected as one of “12 Fascinating Projects From the Bleeding Edge of Interaction Design”[1]. In addition, both projects contributed to a general understanding of various design factors. For example, we learned about the feasibility of direct and indirect input and illustrated the benefits of spatially separating input and output elements. Finally, Section 4.4 presented two concepts which aimed at diversifying user-selected gestures. The projects showed that back-

[1] `http://gizmodo.com/12-fascinating-projects-from-the-bleeding-edge-of-inter-1700656949` – accessed: 2016/08/08.

ground images and presentation effects can have a significant impact on gesture selection. The application of the proposed similarity metric (Section 3.5) indicated that both concepts lead to a more diverse set of gestures. In addition, presentation effects were shown to effectively influence starting cells. Nevertheless, the results also showed that habits are hard to break and many gestures remain predictable.

While the authentication mechanisms were specifically tailored to mobile devices, they can be easily adapted to other authentication scenarios. However, most of the authentication mechanisms require a touchscreen. We argue that public terminals are one possible area of application. For example, *SwiPIN* could be used on touch-based ATMs to make PIN-entry more resistant to observation attacks. In addition, *XSide* can be useful in any context where eyes-free interaction is important and *Marbles* could be useful for digitally secured door openers where smudge attacks might be a critical threat. We argue that the flexibility of use is specifically supported by the fact that all concepts allow encrypted string-based storage of secrets. As a consequence, the concepts fit into most backend infrastructures without modification. Finally, we argue that many aspects of the proposed concepts can be transferred to authentication scenarios where user interaction is usually not touch-based. For example, the *SwiPIN* concept could be implemented for public displays using handheld pointing devices or freehand gestures.

Assistance for the Design and Evaluation of Mobile Authentication Mechanisms

In addition to the tangible results concerning the usability and security of current methods and novel solutions, the thesis provided more general insights into the design and evaluation of mobile authentication mechanisms. As a consequence, the outcome of this thesis can support the design and evaluation process on different levels. Firstly, we pointed out specific but important aspects of the design and the evaluation procedure (e.g., measurement). Secondly, we provided feasible models and metrics which support the assessment of gesture-based authentication concepts. Finally, we presented a general approach to investigate authentication concepts and authentication scenarios in a structured way.

All presented projects contributed general insights into specific aspects of the design and evaluation procedure of mobile authentication methods. For example, Section 4.2 illustrated the importance of correct time measurement and Section 3.2 revealed that unobtrusive experience sampling is a powerful tool to gather insights into the authentication context. Overall, we showed the importance of longitudinal field studies and evaluations outside of the lab. In this regard, Section 4.4 particularly indicated that repeated-measure designs are not suited to assess password selection strategies. In addition, Chapter 4 illustrated the benefits of an iterative user-centered design approach which utilizes both low-fidelity and high-fidelity prototypes. While the individual findings were discussed within the respective sections, Section 5.3 summarized the most important aspects and discussed different study designs.

In addition to the presentation of best practices, we provided concrete tools for the assessment of gesture-based authentication mechanisms. Section 3.3 presented a novel taxonomy for gesture-based input errors. The taxonomy enables the systematic qualitative analysis of logged input errors and revealed that most failures resulted from inaccuracy and not from memorability issues. Section 3.4 provided a prediction model for the observability of grid-based gestures. The model assesses the vulnerability of a given gesture and can be utilized to measure the strength of a gesture concerning observation attacks. Section 3.5 proposed a novel similarity metric which can be used to assess the practical password space of grid-based gestures. Section 4.4 illustrated how the metric can be applied to evaluate the effects of novel security measures. Finally, Section 5.2 provided a more general tool for the development of authentication systems. Even though we make no claim to completeness, we argue that, once the design goals of a project are set, the provided map of interrelations can serve as a valuable resource for the design process. In this regard, we provided two illustrative examples to show how researchers and developers can utilize the findings for both the development of novel solutions and the troubleshooting of existing methods.

While the proposed metrics are more specific to the context of gesture-based authentication, we assume that the observed interconnections of design and problem space are applicable to other authentication scenarios. However, depending on the considered authentication approach and context, the presented model needs to be modified and extended. For instance, the use of hardware keyboards would hamper the applicability of randomly positioned input elements. Nevertheless, we argue that the presented approach can be used as reference framework for other research approaches. Beyond that, we argue that the thesis illustrated a useful problem-solving approach (in terms of design and problem space analysis) which can be adapted to other applied research projects.

6.2 Directions for Future Work

Although the presented research contributed valuable insights into risks and potentials of authentication on mobile devices, it raised new questions and opened opportunities for future research. While each project raised specific issues which are worth investigating, this Section provides a broader perspective of future research directions concerning mobile devices.

Understand Real-World Security Aspects

This thesis illustrated that field studies and controlled experiments are well suited to assess the real-world usability of authentication mechanisms. However, it also illustrated that real-world security is significantly harder to assess. The performed research represents a first step towards quantifying real-world risks and towards understanding security in the wild. However, the presented results had to be based on self-reported data and simulated attacks. We assume that such data can only provide approximations to real world factors. Therefore, we claim that password selection and often discussed practical risks need to be evaluated based on real-world data.

Firstly, we argue that the practical password space of established mechanisms and novel systems should be assessed based on real user-selected secrets. However, since such secrets are usually stored on the device, collecting real-world data represents a major challenge in the mobile context. In addition, researchers should strive for analyzing real-world enrolments, which are embedded in productively used systems, to evaluate password selection strategies. Secondly, we argue that the actual threat of observation attacks and smudge attacks needs to be investigated in the wild. So far, the development of more secure interaction methods has to be justified by the theoretical existence of such risks. However, the number and nature of such attacks is largely unknown. Quantifying the real-world risk would allow to reconsider the requirements of mobile authentication mechanisms.

In summary, we argue that the thorough understanding of real-world threats represents a major challenge of future work. However, gaining this knowledge would allow to develop security mechanisms which are even more tailored to the mobile context and further motivate the investigation of security-optimized concepts.

Enable Context-based Security Mechanisms

Especially Section 3.2 indicated that mobile authentication mechanisms could benefit from context-based access models. We showed that sensitive data is seldom accessed and unlocks often take place in trusted environments. We claim that in such cases, mobile authentication is dispensable and recognizing such dispensable authentication events would help to increase both usability and security. On the one hand, reducing the number of authentications would reduce unnecessary authentication effort. On the other hand, reducing the number of authentications would reduce the potential for observation attacks and smudge attacks. We argue that future research should especially investigate data-dependent, behavioral and environmental cues to provide context-based security mechanisms for the mobile context.

Alternative authentication concepts (e.g., [234]) and implicit authentication (e.g., [137]) have been discussed for several years. However, we argue that the functionality of up-to-date devices and the sensor-rich environments of the Internet of Things (IoT) [110] can finally enable reliable and practical solutions. It has already been shown that most users desire a more flexible control model than all-or-nothing access [124]. Therefore, we argue that the feasibility of data-driven access models should be further explored. One major challenge is the assessment of data sensitivity. User-based assessment is likely to generate a high configuration overhead, automatic assessment is likely to be error-prone. In addition, we assume that the actual sensitivity of the content is depending on the current context. Therefore, we argue that flexible access models need to be combined with other context-based security mechanisms. While such context-based authentication methods [121,206,220] have already been discussed, we argue that the progress of smart environments [110] will open up new possibilities. We claim that future work should consider utilizing information of smart interconnected devices like fridges, vacuum cleaners or light bulbs to effectively enrich the context of the user. Finally, future research should further strive to exploit the sensors of the mobile device itself and investigate the feasibility of behavioral cues (e.g., [69,137,199]) to improve usability and security.

In summary, we argue that a reliable and detailed understanding of the context of mobile devices remains a major challenge. However, the development of novel sensor-rich environments and feature-rich devices opens new opportunities for future research.

Analyze the Interplay between Biometrics and Knowledge-based Solutions

Although we argue that biometric authentication is hardly feasible without knowledge-based fallback mechanisms (see Section 1.1), we assume that biometric solutions will become more popular within the next years. While the reasonable use of biometrics can indeed increase usability and security of mobile devices, it will certainly affect the use and perception of knowledge-based solutions. We assume that the increasing popularity of biometrics will push knowledge-based authentication into the background. On the one hand, knowledge-based authentication will be used less frequently. On the other hand, knowledge-based systems will no longer be perceived as primary authentication mechanisms but rather serve as secondary fallback mechanisms. This paradigm shift is likely to change the requirements of knowledge-based concepts and poses new challenges for future research.

While the primary use of biometric solutions would theoretically enable the usage of stronger fallback mechanisms without significant decreases in usability, Cherapau et al. [50] showed that most users continue using four-digit PINs. We argue that PINs and gestures might even become less secure when used in the context of fallback mechanisms. First of all, such secrets are used less frequently than in the context of primary authentication. Secondly, users may perceive such secrets as less important as they "only" serve as fallback solution. We assume that both aspects are likely to influence password choice and password storage habits. As a consequence, we claim that the interplay between biometrics and knowledge-based solutions needs further investigation. On the one hand, future research should assess the suitability of established concepts. On the other hand, the new demands of such fallback mechanisms might require the development of novel authentication concepts.

In summary, we argue that the increasing popularity of biometric solutions is likely to affect the use and the requirements for knowledge-based solutions. We claim that future research should strive at understanding the interplay of both authentication approaches and consider developing novel solutions which are particularly tailored to the context of fallback authentication. However, we argue that this type of fallback mechanisms significantly differs from traditional fallback solutions which are very rarely needed [114].

Investigate the Strengths and Weaknesses of Novel Types of Mobile Devices

The smartphone is the most prominent representative of mobile devices and wearable devices like smartwatches have just started to become popular[2]. We assume that the number of mobile devices will further grow and other device classes will become available over the next years. We argue that such novel types of mobile devices are likely to have specific requirements for usability and security and will pose new research questions.

[2] `http://www.idc.com/getdoc.jsp?containerId~=~prUS40846515` – last accessed: 2016/08/09

Besides smartphones, smartwatches have already been focused in recent research projects. It has been shown that wearable devices are well-suited to collect behavioral cues for authentication (e.g., [141, 162]). In addition, research came up with novel authentication scenarios (e.g., [93]) and novel authentication concepts (e.g., [291]). However, we assume that touch-based wearable devices are only the beginning of a new interconnected era of mobile devices. Novel areas of research may include authentication on smart glasses [217] or authentication on devices without displays. In addition, future threat models and authentication concepts might consider virtual reality environments. Finally, the interconnection of smart devices is likely to create new security challenges [21]. Research questions include when to authenticate a user and how to propagate access rights between different types of mobile devices.

In summary, we argue that novel types of mobile devices will create new demands on usability and security. While we assume that most of the gathered insights presented in this thesis can be transferred to other touch-based mobile devices, novel form factors may have significantly different requirements. We assume that providing usable and secure security concepts for novel classes of mobile devices will be a major challenge for future research.

6.3 Closing Remarks

This thesis aimed at providing an in-depth understanding of the risks and potentials of graphical and gesture-based authentication for touchscreen mobile devices. By the beginning of this research project in 2012, Android had just recently announced the activation of 190 million devices worldwide[3]. Only four years later, 1.4 billion people were using Android devices[4]. Based on such numbers, I claim that we are just at the beginning of a new era of interconnected devices. Within the next years, the average user will probably utilize smartwatches, smart fridges, smart televisions and smart cars. I assume that the results presented in my thesis will be valuable for some of these devices but certainly not for all. Therefore, I argue that the presented research approach is even more important than the specific findings. With reference to Bruce Schneier, who claimed that "The more technological a society is, the greater the security gap is"[5], I argue that usable security research is likely to become even more important in the future. With this thesis, I want to raise awareness for the importance of a problem-oriented research process. I argue that it is crucial to understand the problem space before proposing novel (security) solutions. Too often, security mechanisms are designed without consideration of the context of use and thus lack practical relevance. The design of novel security concepts should never be an end in itself but strive for having a significant impact on real-world issues. Ultimately, human-centered security mechanisms have to be used to actually provide real-life security.

[3] http://thenextweb.com/google/2011/10/13/google-190-million-android-devices-activated-worldwide-thats-about-576900-a-day-since-may/ – last accessed: 2016/09/08/

[4] http://www.ubergizmo.com/2015/09/over-1-4-billion-people-are-now-using-android/ – last accessed: 2016/09/08/

[5] https://books.google.de/books?id˜=˜WpEfAwAAQBAJ&pg˜=˜PT239&dq – last accessed: 2016/08/09

BIBLIOGRAPHY

[1] A. Adams and M. A. Sasse. Users are not the enemy. *Communications of the ACM*, 42(12), Dec. 1999, pages 40–46.

[2] A. Adams, M. A. Sasse, and P. Lunt. Making passwords secure and usable. In *Proceedings of HCI on People and Computers XII*, HCI 97, London, UK, (1997). Springer-Verlag, pages 1–19.

[3] J. A. Adams. A closed-loop theory of motor learning. *Journal of Motor Behavior*, 3(2), PMID: 15155169, (1971), pages 111–150.

[4] K. Airowaily and M. Alrubaian. Oily residuals security threat on smart phones. In *Robot, Vision and Signal Processing (RVSP), 2011 First International Conference on*. IEEE, (2011), pages 300–302.

[5] F. Alsulaiman and A. Saddik. A novel 3d graphical password schema. In *Virtual Environments, Human-Computer Interfaces and Measurement Systems, Proceedings of 2006 IEEE International Conference on*, July 2006, pages 125–128.

[6] I. Altiok, S. Uellenbeck, and T. Holz. Graphneighbors: Hampering shoulder-surfing attacks on smartphones. In *Sicherheit*, (2014), pages 25–35.

[7] M. D. Amruth and K. Praveen. Android smudge attack prevention techniques. In *Intelligent Systems Technologies and Applications: Volume 2*, Cham, (2016). Springer International Publishing, pages 23–31.

[8] J. R. Anderson. *Cognitive psychology and its implications.* WH Freeman/Times Books/Henry Holt & Co, (1990).

[9] S. Andrews, D. A. Ellis, H. Shaw, and L. Piwek. Beyond self-report: tools to compare estimated and real-world smartphone use. *PloS one*, 10(10), (2015), pages 1–9.

[10] P. Andriotis, T. Tryfonas, and G. Oikonomou. Complexity metrics and user strength perceptions of the pattern-lock graphical authentication method. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*, volume 8533 of *Lecture Notes in Computer Science*. Springer International Publishing, (2014), pages 115–126.

[11] P. Andriotis, T. Tryfonas, G. Oikonomou, and C. Yildiz. A pilot study on the security of pattern screen-lock methods and soft side channel attacks. In *Proceedings of the Sixth ACM Conference on Security and Privacy in Wireless and Mobile Networks*, WiSec '13, New York, NY, USA, (2013). ACM, pages 1–6.

[12] J. Angulo and E. Wästlund. Exploring touch-screen biometrics for user identification on smart phones. In *Privacy and Identity Management for Life*, volume 375 of *IFIP Advances in Information and Communication Technology*. Springer Berlin Heidelberg, (2012), pages 130–143.

[13] U. Ansorge, M. Heumann, and I. Scharlau. Influences of visibility, intentions, and probability in a peripheral cuing task. *Consciousness and Cognition*, 11(4), (2002), pages 528–545.

[14] M. Anwar and A. Imran. A comparative study of graphical and alphanumeric passwords for mobile device authentication. In *Proceedings of the 26th Modern AI and Cognitive Science Conference 2015*, Greensboro, NC, USA, April 25-26, 2015. 2015, pages 13–18.

[15] M. Arianezhad, D. Stebila, and B. Mozaffari. Usability and security of gaze-based graphical grid passwords. In *Financial Cryptography and Data Security*, volume 7862 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, (2013), pages 17–33.

[16] P. Arnold. An analysis of graphical password selection, strength, and entropy. Bachelor thesis, Ludwig-Maximilians-Universität München, (2014).

[17] R. C. Atkinson and R. M. Shiffrin. Human Memory: A Proposed System and Its Control Processes. *The Psychology of Learning and Motivation: Advances in Research and Theory*, Vol. 2, (1968), pages 89 – 195.

[18] A. J. Aviv, K. Gibson, E. Mossop, M. Blaze, and J. M. Smith. Smudge attacks on smartphone touch screens. In *Proceedings of the 4th USENIX Conference on Offensive Technologies*, WOOT'10, Berkeley, CA, USA, (2010). USENIX Association, pages 1–7.

[19] A. J. Aviv, B. Sapp, M. Blaze, and J. M. Smith. Practicality of accelerometer side channels on smartphones. In *Proceedings of the 28th Annual Computer Security Applications Conference*, ACSAC '12, New York, NY, USA, (2012). ACM, pages 41–50.

[20] S. Azenkot, K. Rector, R. Ladner, and J. Wobbrock. Passchords: Secure multi-touch authentication for blind people. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '12, New York, NY, USA, (2012). ACM, pages 159–166.

[21] S. Babar, P. Mahalle, A. Stango, N. Prasad, and R. Prasad. Proposed security model and threat taxonomy for the internet of things (iot). In *Recent Trends in Network Security and Applications: Third International Conference, CNSA 2010, Chennai, India, July 23-25, 2010. Proceedings*, Berlin, Heidelberg, (2010). Springer Berlin Heidelberg, pages 420–429.

[22] P. Baudisch and G. Chu. Back-of-device interaction allows creating very small touch devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, New York, NY, USA, (2009). ACM, pages 1923–1932.

[23] A. Beautement, M. A. Sasse, and M. Wonham. The compliance budget: Managing security behaviour in organisations. In *Proceedings of the 2008 Workshop on New Security Paradigms*, NSPW '08, New York, NY, USA, (2008). ACM, pages 47–58.

[24] M. Becher, F. Freiling, J. Hoffmann, T. Holz, S. Uellenbeck, and C. Wolf. Mobile security catching up? revealing the nuts and bolts of the security of mobile devices. In *2011 IEEE Symposium on Security and Privacy (SP)*, May 2011, pages 96–111.

[25] E. Bernat, S. Bunce, and H. Shevrin. Event-related brain potentials differentiate positive and negative mood adjectives during both supraliminal and subliminal visual processing. *International Journal of Psychophysiology*, 42(1), (2001), pages 11–34.

[26] A. Bianchi and I. Oakley. Multiplexed input to protect against casual observers. In *Proceedings of HCI Korea*, HCIK '15, South Korea, (2014). Hanbit Media, Inc., pages 7–11.

[27] A. Bianchi, I. Oakley, V. Kostakos, and D. S. Kwon. The phone lock: Audio and haptic shoulder-surfing resistant pin entry methods for mobile devices. In *Proceedings of the Fifth International Conference on Tangible, Embedded, and Embodied Interaction*, TEI '11, New York, NY, USA, (2011). ACM, pages 197–200.

[28] A. Bianchi, I. Oakley, and D. Kwon. Spinlock: a single-cue haptic and audio pin input technique for authentication. In *International Workshop on Haptic and Audio Interaction Design*, Berlin, Heidelberg, (2011). Springer Berlin Heidelberg, pages 81–90.

[29] A. Bianchi, I. Oakley, and D. S. Kwon. Counting clicks and beeps: Exploring numerosity based haptic and audio pin entry. *Interacting with Computers*, 24(5), (2012), pages 409 – 422.

[30] A. Bianchi, I. Oakley, and D.-S. Kwon. Open sesame: Design guidelines for invisible passwords. *Computer*, 45(4), April 2012, pages 58–65.

[31] K. Bicakci, N. Atalay, M. Yuceel, H. Gurbaslar, and B. Erdeniz. Towards usable solutions to graphical password hotspot problem. In *Computer Software and Applications Conference, 2009. COMPSAC '09. 33rd Annual IEEE International*, volume 2, July 2009, pages 318–323.

[32] R. Biddle, S. Chiasson, and P. Van Oorschot. Graphical passwords: Learning from the first twelve years. *ACM Computing Surveys (CSUR)*, 44(4), Sept. 2012, pages 19:1–19:41.

[33] M. Bishop and D. V. Klein. Improving system security via proactive password checking. *Computers & Security*, 14(3), (1995), pages 233 – 249.

[34] R. A. Block and R. P. Gruber. Time perception, attention, and memory: A selective review. *Acta Psychologica*, 149, Including Special section articles of Temporal Processing Within and Across Senses - Part-2, (2014), pages 129 – 133.

[35] G. E. Blonder. Graphical password, Sept. 24 1996. US Patent 5,559,961.

[36] H. Bojinov and D. Boneh. Mobile token-based authentication on a budget. In *Proceedings of the 12th Workshop on Mobile Computing Systems and Applications*, HotMobile '11, New York, NY, USA, (2011). ACM, pages 14–19.

[37] J. Bonneau. *Guessing human-chosen secrets*. PhD thesis, University of Cambridge, (2012).

[38] J. Bonneau. The science of guessing: Analyzing an anonymized corpus of 70 million passwords. In *2012 IEEE Symposium on Security and Privacy*, May 2012, pages 538–552.

[39] J. Bonneau, C. Herley, P. van Oorschot, and F. Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *2012 IEEE Symposium on Security and Privacy (SP)*, May 2012, pages 553–567.

[40] J. Bonneau and S. Schechter. Towards reliable storage of 56-bit secrets in human memory. In *Proceedings of the 23rd USENIX Conference on Security Symposium*, SEC'14, Berkeley, CA, USA, (2014). USENIX Association, pages 607–623.

[41] S. Brostoff and M. A. Sasse. Are passfaces more usable than passwords? a field trial investigation. In *People and Computers XIV — Usability or Else!: Proceedings of HCI 2000*, London, (2000). Springer London, pages 405–424.

[42] A. S. Brown, E. Bracken, S. Zoccoli, and K. Douglas. Generating and remembering passwords. *Applied Cognitive Psychology*, 18(6), (2004), pages 641–651.

[43] B. Brunkow. Using gestures to protect pin-entry from shoulder surfing. Bachelor thesis, Ludwig-Maximilians-Universität München, (2014).

[44] K. Bryant and J. Campbell. User behaviours associated with password security and management. *Australasian Journal of Information Systems*, 14(1), (2006), pages 81–100.

[45] A. Bulling, F. Alt, and A. Schmidt. Increasing the security of gaze-based cued-recall graphical passwords using saliency masks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, New York, NY, USA, (2012). ACM, pages 3011–3020.

[46] J. Bunnell, J. Podd, R. Henderson, R. Napier, and J. Kennedy-Moffat. Cognitive, associative and conventional passwords: Recall and guessing rates. *Computers & Security*, 16(7), (1997), pages 629 – 641.

[47] W. E. Burr, D. F. Dodson, E. M. Newton, R. A. Perlner, W. T. Polk, S. Gupta, and E. A. Nabbus. Sp 800-63-1. electronic authentication guideline. Technical report, NIST, Gaithersburg, MD, United States, (2011).

[48] A. Busch. Designing a graphical shoulder surfing resistant authentication mechanism for off-the-shelf smartphones. Bachelor thesis, Ludwig-Maximilians-Universität München, (2013).

[49] J. Cederberg. *A course in modern geometries*. Springer Science & Business Media, (2013).

[50] I. Cherapau, I. Muslukhov, N. Asanka, and K. Beznosov. On the impact of touch id on iphone passcodes. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, Ottawa, July 2015. USENIX Association, pages 257–276.

[51] M. Cherubini and N. Oliver. A Refined Experience Sampling Method to Capture Mobile User Experience. In *International Workshop of Mobile User Experience Research part of CHI'2009*, (2009), pages 1–12.

[52] H.-Y. Chiang and S. Chiasson. Improving user authentication on mobile devices: A touchscreen graphical password. In *Proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services*, MobileHCI '13, New York, NY, USA, (2013). ACM, pages 251–260.

[53] S. Chiasson, R. Biddle, and P. C. van Oorschot. A second look at the usability of click-based graphical passwords. In *Proceedings of the 3rd Symposium on Usable Privacy and Security*, SOUPS '07, New York, NY, USA, (2007). ACM, pages 1–12.

[54] S. Chiasson, A. Forget, R. Biddle, and P. van Oorschot. User interface design affects security: patterns in click-based graphical passwords. *International Journal of Information Security*, 8(6), (2009), pages 387–398.

[55] S. Chiasson, A. Forget, E. Stobert, P. C. van Oorschot, and R. Biddle. Multiple password interference in text passwords and click-based graphical passwords. In *Proceedings of the 16th ACM Conference on Computer and Communications Security*, CCS '09, New York, NY, USA, (2009). ACM, pages 500–511.

[56] S. Chiasson, E. Stobert, A. Forget, R. Biddle, and P. Van Oorschot. Persuasive cued click-points: Design, implementation, and evaluation of a knowledge-based authentication mechanism. *Dependable and Secure Computing, IEEE Transactions on*, 9(2), March 2012, pages 222–235.

[57] S. Chiasson, P. C. van Oorschot, and R. Biddle. Graphical password authentication using cued click points. In *Proceedings of the 12th European Symposium On Research In Computer Security*, ESORICS 2007, Berlin, Heidelberg, (2007). Springer Berlin Heidelberg, pages 359–374.

[58] E. Chin, A. P. Felt, V. Sekar, and D. Wagner. Measuring user confidence in smartphone security and privacy. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, New York, NY, USA, (2012). ACM, pages 1:1–1:16.

[59] M. K. Chong and G. Marsden. Exploring the use of discrete gestures for authentication. In T. Gross, J. Gulliksen, P. Kotzé, L. Oestreicher, P. Palanque, R. O. Prates, and M. Winckler (editors). *Human-Computer Interaction – INTERACT 2009: 12th IFIP TC 13 International Conference, Uppsala, Sweden, August 24-28, 2009, Proceedings, Part II*, Berlin, Heidelberg, (2009). Springer Berlin Heidelberg, pages 205–213.

[60] V. Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3), (1979), pages 233–235.

[61] J. Citty and D. R. Hutchings. Tapi: Touch-screen authentication using partitioned images. Technical report, Elon University, (2010).

[62] F. J. Corbató, M. Merwin-Daggett, and R. C. Daley. An experimental time-sharing system. In *Proceedings of the May 1-3, 1962, Spring Joint Computer Conference*, AIEE-IRE '62 (Spring), New York, NY, USA, (1962). ACM, pages 335–344.

[63] F. J. Corbató, J. H. Saltzer, and C. T. Clingen. Multics: The first seven years. In *Proceedings of the May 16-18, 1972, Spring Joint Computer Conference*, AFIPS '72 (Spring), New York, NY, USA, (1972). ACM, pages 571–583.

[64] D. Damopoulos, G. Kambourakis, and S. Gritzalis. From keyloggers to touchloggers: Take the rough with the smooth. *Computers & Security*, 32, (2013), pages 102 – 114.

[65] A. Das, J. Bonneau, M. Caesar, N. Borisov, and X. Wang. The tangled web of password reuse. In *Proceedings of The Network and Distributed System Security Symposium*, volume 14 of *NDSS'14*. Internet Society, (2014), pages 1–15.

[66] D. Davis, F. Monrose, and M. K. Reiter. On user choice in graphical password schemes. In *SSYM'04: Proceedings of the 13th conference on USENIX Security Symposium*, Berkeley, CA, USA, (2004). USENIX Association, pages 11–11.

[67] A. De Angeli, M. Coutts, L. Coventry, G. I. Johnson, D. Cameron, and M. H. Fischer. Vip: a visual approach to user authentication. In *AVI '02: Proceedings of the Working Conference on Advanced Visual Interfaces*, New York, NY, USA, (2002). ACM, pages 316–323.

[68] A. De Angeli, L. Coventry, G. Johnson, and K. Renaud. Is a picture really worth a thousand words? exploring the feasibility of graphical authentication systems. *International Journal of Human-Computer Studies*, 63(1), (2005), pages 128–152.

[69] A. De Luca, A. Hang, F. Brudy, C. Lindner, and H. Hussmann. Touch me once and i know it's you!: Implicit authentication based on touch screen patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, New York, NY, USA, (2012). ACM, pages 987–996.

[70] A. De Luca, A. Hang, E. von Zezschwitz, and H. Hussmann. I feel like i'm taking selfies all day!: Towards understanding biometric authentication on smartphones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, New York, NY, USA, (2015). ACM, pages 1411–1414.

[71] A. De Luca, M. Harbach, E. von Zezschwitz, M.-E. Maurer, B. E. Slawik, H. Hussmann, and M. Smith. Now you see me, now you don't: Protecting smartphone authentication from shoulder surfers. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, New York, NY, USA, (2014). ACM, pages 2937–2946.

[72] A. De Luca, K. Hertzschuch, and H. Hussmann. Colorpin: Securing pin entry through indirect input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, New York, NY, USA, (2010). ACM, pages 1103–1106.

[73] A. De Luca, E. Von Zezschwitz, and H. Hußmann. Vibrapass: secure authentication based on shared lies. In *Proceedings of the 27th international conference on Human factors in computing systems*. ACM, (2009), pages 913–916.

[74] A. De Luca, E. von Zezschwitz, N. D. H. Nguyen, M.-E. Maurer, E. Rubegni, M. P. Scipioni, and M. Langheinrich. Back-of-device authentication on smartphones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, New York, NY, USA, (2013). ACM, pages 2389–2398.

[75] A. De Luca, E. von Zezschwitz, L. Pichler, and H. Hussmann. Using fake cursors to secure on-screen password entry. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, New York, NY, USA, (2013). ACM, pages 2399–2402.

[76] M. Dell'Amico, P. Michiardi, and Y. Roudier. Password strength: an empirical analysis. In *Proceedings of the 29th conference on Information communications*, INFOCOM'10, Piscataway, NJ, USA, (2010). IEEE Press, pages 983–991.

[77] R. Dhamija and A. Perrig. Déjà vu: a user study using images for authentication. In *Proceedings of the 9th conference on USENIX Security Symposium*, SSYM'00, Berkeley, CA, USA, (2000). USENIX Association, pages 45–58.

[78] A. E. Dirik, N. Memon, and J.-C. Birget. Modeling user choice in the passpoints graphical password scheme. In *Proceedings of the 3rd symposium on Usable privacy and security*. ACM, (2007), pages 20–28.

[79] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor. Are your participants gaming the system?: Screening mechanical turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, New York, NY, USA, (2010). ACM, pages 2399–2402.

[80] P. Dunphy, A. P. Heiner, and N. Asokan. A closer look at recognition-based graphical passwords on mobile devices. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, New York, NY, USA, (2010). ACM, pages 3:1–3:12.

[81] P. Dunphy, J. Nicholson, and P. Olivier. Securing passfaces for description. In *Proceedings of the 4th Symposium on Usable Privacy and Security*, SOUPS '08, New York, NY, USA, (2008). ACM, pages 24–35.

[82] P. Dunphy and P. Olivier. On automated image choice for secure and usable graphical passwords. In *Proceedings of the 28th Annual Computer Security Applications Conference*, ACSAC '12, New York, NY, USA, (2012). ACM, pages 99–108.

[83] P. Dunphy and J. Yan. Do background images improve draw a secret graphical passwords? In *Proceedings of the 14th ACM conference on Computer and communications security*. ACM, (2007), pages 36–47.

[84] S. Egelman, S. Jain, R. S. Portnoff, K. Liao, S. Consolvo, and D. Wagner. Are you ready to lock? In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, New York, NY, USA, (2014). ACM, pages 750–761.

[85] S. Egelman, S. Jain, R. S. Portnoff, K. Liao, S. Consolvo, and D. Wagner. Are you ready to lock? In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, New York, NY, USA, (2014). ACM, pages 750–761.

[86] S. Egelman and E. Peer. Scaling the security wall: Developing a security behavior intentions scale (sebis). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, New York, NY, USA, (2015). ACM, pages 2873–2882.

[87] S. Egelman, A. Sotirakopoulos, I. Muslukhov, K. Beznosov, and C. Herley. Does my password go up to eleven?: The impact of password meters on password selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, New York, NY, USA, (2013). ACM, pages 2379–2388.

[88] M. Eiband. The influence of background images on android (un)lock pattern selection. Master thesis, Ludwig-Maximilians-Universität München, (2015).

[89] H. Ellis. The science behind passfaces. Technical report, passfaces tm, (2004).

[90] K. M. Everitt, T. Bragin, J. Fogarty, and T. Kohno. A comprehensive study of frequency, interference, and training of multiple graphical passwords. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, New York, NY, USA, (2009). ACM, pages 889–898.

[91] D. Fallman. Why research-oriented design isn't design-oriented research: On the tensions between design and research in an implicit design discipline. *Knowledge, Technology & Policy*, 20(3), (2007), pages 193–200.

[92] A. P. Felt, S. Egelman, and D. Wagner. I've got 99 problems, but vibration ain't one: A survey of smartphone users' concerns. In *Proceedings of the Second ACM Workshop on Security and Privacy in Smartphones and Mobile Devices*, SPSM '12, New York, NY, USA, (2012). ACM, pages 33–44.

[93] A. Ferrari, D. Puccinelli, and S. Giordano. Gesture-based soft authentication. In *11th International Conference on Wireless and Mobile Computing, Networking and Communications*, WiMob'15. IEEE, Oct 2015, pages 771–777.

[94] E. Fleishman and J. Parker Jr. Factors in the retention and relearning of perceptual-motor skill. *Journal of Experimental Psychology*, 64(3), (1962), pages 215–226.

[95] D. Florencio and C. Herley. A large-scale study of web password habits. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, New York, NY, USA, (2007). ACM, pages 657–666.

[96] D. Florêncio and C. Herley. Where do security policies come from? In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, New York, NY, USA, (2010). ACM, pages 10:1–10:14.

[97] D. Florêncio, C. Herley, and P. C. Van Oorschot. An administrator's guide to internet password research. In *Proceedings of the 28th USENIX Conference on Large Installation System Administration*, LISA'14, Berkeley, CA, USA, (2014). USENIX Association, pages 35–52.

[98] D. Florêncio, C. Herley, and P. C. Van Oorschot. Password portfolios and the finite-effort user: Sustainably managing large numbers of accounts. In *Proceedings of the 23rd USENIX Conference on Security Symposium*, SEC'14, Berkeley, CA, USA, (2014). USENIX Association, pages 575–590.

[99] A. Forget, S. Chiasson, and R. Biddle. Persuasion as education for computer security. Technical Report 1, Carleton University, (2007).

[100] A. Forget, S. Chiasson, and R. Biddle. Shoulder-surfing resistance with eye-gaze entry in cued-recall graphical passwords. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, New York, NY, USA, (2010). ACM, pages 1107–1110.

[101] P. Fraisse. Perception and estimation of time. *Annual review of psychology*, 35(1), (1984), pages 1–37.

[102] M. A. Gallagher and M. D. Byrne. Modeling password entry on a mobile device. In *Proceedings of the International Conference on Cognitive Modeling*, ICCM'15, (2015), pages 45–50.

[103] H. Gao, X. Guo, X. Chen, L. Wang, and X. Liu. Yagp: Yet another graphical password strategy. In *In Proceedings of the Computer Security Applications Conference*, ACSAC'08. IEEE, (2008), pages 121–129.

[104] H. Gao, W. Jia, F. Ye, and L. Ma. A survey on the use of graphical passwords in security. *Journal of Software*, 8(7), (2013), pages 1678–1698.

[105] H. Gao, Z. Ren, X. Chang, X. Liu, and U. Aickelin. A new graphical password scheme resistant to shoulder-surfing. In *International Conference on Cyberworlds*, CW'10, Oct 2010, pages 194–199.

[106] S. Garfinkel and H. R. Lipford. Usable security: History, themes, and challenges. *Synthesis Lectures on Information Security, Privacy, and Trust*, 5(2), (2014), pages 1–124.

[107] S. Gaw and E. W. Felten. Password management strategies for online accounts. In *Proceedings of the Second Symposium on Usable Privacy and Security*, SOUPS '06, New York, NY, USA, (2006). ACM, pages 44–55.

[108] V. Goel and P. Pirolli. The structure of design problem spaces. *Cognitive science*, 16(3), (1992), pages 395–429.

[109] B. Grawemeyer and H. Johnson. Using and managing multiple passwords: A week to a view. *Interacting with Computers*, 23(3), (2011), pages 256–267.

[110] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami. Internet of things (iot): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, 29(7), (2013), pages 1645 – 1660.

[111] J. Guerra-Casanova, C. Sánchez-Ávila, G. Bailador, and A. de Santos Sierra. Authentication in mobile devices through hand gesture recognition. *International Journal of Information Security*, 11(2), (2012), pages 65–83.

[112] J. Gugenheimer, A. De Luca, H. Hess, S. Karg, D. Wolf, and E. Rukzio. Colorsnakes: Using colored decoys to secure authentication in sensitive contexts. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '15, New York, NY, USA, (2015). ACM, pages 274–283.

[113] M. Hafiz, A. Abdullah, N. Ithnin, and H. Mammi. Towards identifying usability and security features of graphical password in knowledge based authentication technique. In *Second Asia International Conference on Modeling Simulation*, AICMS'08, May 2008, pages 396–403.

[114] A. Hang, A. De Luca, E. von Zezschwitz, M. Demmler, and H. Hussmann. Locked your phone? buy a new one? from tales of fallback authentication on smartphones to actual concepts. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '15, New York, NY, USA, (2015). ACM, pages 295–305.

[115] A. Hang, E. Von Zezschwitz, A. De Luca, and H. Hussmann. Too much information!: user attitudes towards smartphone sharing. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*. ACM, (2012), pages 284–287.

[116] S. T. Haque, M. Wright, and S. Scielzo. Hierarchy of users' web passwords: Perceptions, practices and susceptibilities. *International Journal of Human-Computer Studies*, 72(12), (2014), pages 860 – 874.

[117] M. Harbach, A. De Luca, N. Malkin, and S. Egelman. Keep on lockin' in the free world: A multi-national comparison of smartphone locking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, New York, NY, USA, (2016). ACM, pages 4823–4827.

[118] M. Harbach, E. von Zezschwitz, A. Fichtner, A. De Luca, and M. Smith. It's a hard lock life: A field study of smartphone (un)locking behavior and risk perception. In *Symposium On Usable Privacy and Security*, SOUPS'14, Menlo Park, CA, July 2014. USENIX Association, pages 213–230.

[119] K. Hasan, X.-D. Yang, H.-N. Liang, and P. Irani. How to position the cursor?: An exploration of absolute and relative cursor positioning for back-of-device input. In *Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services*, MobileHCI '12, New York, NY, USA, (2012). ACM, pages 103–112.

[120] M. Hasegawa, N. Isogai, and S. Kato. On design of audio instructions for multisensory authentication for portable touchscreen device. In *Poster at the Symposium on Usable Privacy and Security*, SOUPS'12. CMU CyLab, (2012), pages 1–2.

[121] E. Hayashi, S. Das, S. Amini, J. Hong, and I. Oakley. Casa: Context-aware scalable authentication. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, New York, NY, USA, (2013). ACM, pages 3:1–3:10.

[122] E. Hayashi, R. Dhamija, N. Christin, and A. Perrig. Use your illusion: secure authentication usable anywhere. In *Proceedings of the 4th symposium on Usable privacy and security*, SOUPS '08, New York, NY, USA, (2008). ACM, pages 35–45.

[123] E. Hayashi, M. Maas, and J. I. Hong. Wave to me: User identification using body lengths and natural gestures. In *Proceedings of the SIGCHI Conference on Human*

Factors in Computing Systems, CHI '14, New York, NY, USA, (2014). ACM, pages 3453–3462.

[124] E. Hayashi, O. Riva, K. Strauss, A. J. B. Brush, and S. Schechter. Goldilocks and the two mobile devices: Going beyond all-or-nothing access to a device's applications. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, New York, NY, USA, (2012). ACM, pages 2:1–2:11.

[125] K. Helkala, N. Svendsen, P. Thorsheim, and A. Wiehe. Cracking associative passwords. In *Secure IT Systems*, volume 7617 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, (2012), pages 153–168.

[126] C. Herley. So long, and no thanks for the externalities: The rational rejection of security advice by users. In *Proceedings of the 2009 Workshop on New Security Paradigms Workshop*, NSPW '09, New York, NY, USA, (2009). ACM, pages 133–144.

[127] C. Herley and D. Florencio. How to login from an internet café without worrying about keyloggers. In *Poster at the Symposium on Usable Privacy and Security*, Soups'06, (2006), pages 1–2.

[128] C. Herley, P. C. Oorschot, and A. S. Patrick. Passwords: If we're so smart, why are we still using them? In *Financial Cryptography and Data Security*. Springer-Verlag, Berlin, Heidelberg, (2009), pages 230–237.

[129] C. Herley and P. C. van Oorschot. A research agenda acknowledging the persistence of passwords. *IEEE Security & Privacy*, 10(1), (2012), pages 28–36.

[130] P. F. Ho, Y. H.-S. Kam, M. C. Wee, Y. N. Chong, and L. Y. Por. Preventing shoulder-surfing attack with the concept of concealing the password objects' information. *The Scientific World Journal*, 2014, (2014), pages 1–12.

[131] L. Holt. Increasing real-world security of user ids and passwords. In *Proceedings of the 2011 Information Security Curriculum Development Conference*, InfoSecCD '11, New York, NY, USA, (2011). ACM, pages 34–41.

[132] P. Hoonakker, N. Bornoe, and P. Carayon. Password authentication from a human factors perspective: Results of a survey among end-users. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 53(6), (2009), pages 459–463.

[133] P. G. Inglesant and M. A. Sasse. The true cost of unusable password policies: Password use in the wild. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, New York, NY, USA, (2010). ACM, pages 383–392.

[134] ISO. Iso 9564-1:2011 financial services – personal identification number (pin) management and security – part 1: Basic principles and requirements for pins in card-based systems. Standard, TC 68 Financial services, FEB 2011.

[135] B. Ives, K. R. Walsh, and H. Schneider. The domino effect of password reuse. *Communications of the ACM*, 47(4), Apr. 2004, pages 75–78.

[136] M. Jakobsson and M. Dhiman. The benefits of understanding passwords. In *Mobile Authentication*, SpringerBriefs in Computer Science. Springer New York, (2013), pages 5–24.

[137] M. Jakobsson, E. Shi, P. Golle, and R. Chow. Implicit authentication for mobile devices. In *Proceedings of the 4th USENIX conference on Hot topics in security*, HotSec'09, Berkeley, CA, USA, (2009). USENIX Association, pages 1–6.

[138] W. Jansen. Authenticating mobile device users through image selection. *The Internet Society: Advances in Learning, Commerce and Security*, 1, (2004), pages 183–194.

[139] P. Janssen. On the impact of pattern composition strategies on shoulder-surfing vulnerability. Bachelor thesis, Ludwig-Maximilians-Universität München, (2014).

[140] I. Jermyn, A. Mayer, F. Monrose, M. K. Reiter, and A. D. Rubin. The design and analysis of graphical passwords. In *Proceedings of the 8th Conference on USENIX Security Symposium - Volume 8*, SSYM'99, Berkeley, CA, USA, (1999). USENIX Association, pages 1–14.

[141] A. H. Johnston and G. M. Weiss. Smartwatch-based biometric gait recognition. In *7th International Conference on Biometrics Theory, Applications and Systems*, BTAS'15. IEEE, Sept 2015, pages 1–6.

[142] A. K. Karlson, A. B. Brush, and S. Schechter. Can i borrow your phone?: Understanding concerns when sharing mobile phones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, New York, NY, USA, (2009). ACM, pages 1647–1650.

[143] J. J. Kaye. Self-reported password sharing strategies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, New York, NY, USA, (2011). ACM, pages 2619–2622.

[144] A. Kehr. Memorability and usability of randomized graphical authentication systems for mobile devices. Bachelor thesis, Ludwig-Maximilians-Universität München, (2014).

[145] P. Kelley, S. Komanduri, M. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L. Cranor, and J. Lopez. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *2012 IEEE Symposium on Security and Privacy*, SP'12. IEEE, May 2012, pages 523–537.

[146] W. Z. Khan, M. Y. Aalsalem, and Y. Xiang. A graphical password based system for small mobile devices. *IJCSI International Journal of Computer Science Issues*, 8(2), (2011), pages 145–154.

[147] R. A. Khot, P. Kumaraguru, and K. Srinathan. Wyswye: Shoulder surfing defense for recognition based graphical passwords. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*, OzCHI '12, New York, NY, USA, (2012). ACM, pages 285–294.

[148] A. Kienle. The presentation effect on grid-based passwords. Bachelor thesis, Ludwig-Maximilians-Universität München, (2015).

[149] A. Kienle. The presentation effect on grid-based passwords revisited. Practical research course, Ludwig-Maximilians-Universität Münchenn, (2015).

[150] A. Kiesel. Unbewusste Wahrnehmung. *Psychologische Rundschau*, 60(4), (2009), pages 215–228.

[151] D. Kim, P. Dunphy, P. Briggs, J. Hook, J. Nicholson, J. Nicholson, and P. Olivier. Multi-touch authentication on tabletops. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI'10, New York, NY, USA, (2010). ACM, pages 1093–1102.

[152] Y. Kita, F. Sugai, M. Park, and N. Okazaki. Proposal and its evaluation of a shoulder-surfing attack resistant authentication method: Secret tap with double shift. *International Journal of Cyber-Security and Digital Forensics (IJCSDF)*, 2(1), (2013), pages 48–55.

[153] J. Kjeldskov and J. Paay. A longitudinal review of mobile hci research methods. In *Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services*, MobileHCI '12, New York, NY, USA, (2012). ACM, pages 69–78.

[154] S. T. Klapp and B. W. Haas. Nonconscious influence of masked stimuli on response selection is limited to concrete stimulus-response associations. *Journal of Experimental Psychology: Human Perception and Performance*, 31(1), (2005), pages 193–209.

[155] S. Komanduri, R. Shay, P. G. Kelley, M. L. Mazurek, L. Bauer, N. Christin, L. F. Cranor, and S. Egelman. Of passwords and people: Measuring the effect of password-composition policies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, New York, NY, USA, (2011). ACM, pages 2595–2604.

[156] B. Korte and J. Vygen. *Combinatorial Optimization: Theory and Algorithms*, volume 21 of *Algorithms and Combinatorics*. Springer-Verlag Berlin Heidelberg, 5th edition, (2012).

[157] A. Koslow. A pattern-based authentication method resilient against smudge attacks. Bachelor thesis, Ludwig-Maximilians-Universität München, (2012).

[158] M. Kosugi, T. Suzuki, O. Uchida, and H. Kikuchi. Swipass: Image-based user authentication for touch screen devices. *Journal of Information Processing*, 24(2), (2016), pages 227–236.

[159] C. Kuo, S. Romanosky, and L. F. Cranor. Human selection of mnemonic phrase-based passwords. In *Proceedings of the Second Symposium on Usable Privacy and Security*, SOUPS '06, New York, NY, USA, (2006). ACM, pages 67–78.

[160] T. Kuribara, B. Shizuki, and J. Tanaka. Vibrainput: Two-step pin entry system based on vibration and visual information. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, New York, NY, USA, (2014). ACM, pages 2473–2478.

[161] T. Kwon and S. Na. Tinylock: Affordable defense against smudge attacks on smartphone pattern lock systems. *Computers & Security*, 42(0), (2014), pages 137 – 150.

[162] W.-H. Lee and R. Lee. Implicit sensor-based authentication of smartphone users with smartwatch. In *Proceedings of the Hardware and Architectural Support for Security and Privacy 2016*, HASP 2016, New York, NY, USA, (2016). ACM, pages 9:1–9:8.

[163] M. Li and P. M. Vitnyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer New York, 3 edition, (2008).

[164] D. Lin, P. Dunphy, P. Olivier, and J. Yan. Graphical passwords & qualitative spatial relations. In *Proceedings of the 3rd Symposium on Usable Privacy and Security*, SOUPS '07, New York, NY, USA, (2007). ACM, pages 161–162.

[165] J. Liu, L. Zhong, J. Wickramasuriya, and V. Vasudevan. uwave: Accelerometer-based personalized gesture recognition and its applications. *Pervasive and Mobile Computing*, 5(6), PerCom 2009, (2009), pages 657 – 675.

[166] M. Löchtefeld, C. Hirtz, and S. Gehring. Evaluation of hybrid front- and back-of-device interaction on mobile devices. In *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia*, MUM '13, New York, NY, USA, (2013). ACM, pages 17:1–17:4.

[167] W. Ma, J. Campbell, D. Tran, and D. Kleeman. Password entropy and password quality. In *4th International Conference on Network and System Security*, Sept 2010, pages 583–587.

[168] W. E. Mackay and A.-L. Fayard. Hci, natural science and design: a framework for triangulation across disciplines. In *Proceedings of the 2nd conference on Designing interactive systems: processes, practices, methods, and techniques*. ACM, (1997), pages 223–234.

[169] B. Malek, M. Orozco, and A. El Saddik. Novel shoulder-surfing resistant haptic-based graphical password. In *EuroHaptics 2006*. EuroHaptics Society, July 2006, pages 1–6.

[170] D. Malone and K. Maher. Investigating the distribution of password choices. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, New York, NY, USA, (2012). ACM, pages 301–310.

[171] M. Mannan and P. Van Oorschot. Passwords for both mobile and desktop computers: Obpwd for firefox and android. *USENIX; login*, 37(4), (2012), pages 28–37.

[172] I. Maurer and S. Oberhuber. Analysis of android unlock pattern. Practical research course, Ludwig-Maximilians-Universität Münchenn, (2014).

[173] M.-E. Maurer, R. Waxenberger, and D. Hausen. Broauth: Evaluating different levels of visual feedback for 3d gesture-based authentication. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, New York, NY, USA, (2012). ACM, pages 737–740.

[174] S. Maydebura, D. H. Jeong, and B. Yu. Understanding environmental influences on performing password-based mobile authentication. In *14th International Conference on Information Reuse and Integration*, IRI'13. IEEE, Aug 2013, pages 728–731.

[175] R. E. Mayer and V. K. Sims. For whom is a picture worth a thousand words? extensions of a dual-coding theory of multimedia learning. *Journal of Educational Psychology*, 86(3), (1994), pages 389–401.

[176] R. Mayrhofer and H. Gellersen. Shake well before use: Intuitive and secure pairing of mobile devices. *Transactions on Mobile Computing*, 8(6), June 2009, pages 792–806.

[177] M. L. Mazurek, S. Komanduri, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, P. G. Kelley, R. Shay, and B. Ur. Measuring password guessability for an entire university. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, CCS '13, New York, NY, USA, (2013). ACM, pages 173–186.

[178] W. Melicher, D. Kurilova, S. M. Segreti, P. Kalvani, R. Shay, B. Ur, L. Bauer, N. Christin, L. F. Cranor, and M. L. Mazurek. Usability and security of text passwords on mobile devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, New York, NY, USA, (2016). ACM, pages 527–539.

[179] Y. Meng. Designing click-draw based graphical password scheme for better authentication. In *7th International Conference on Networking, Architecture and Storage*, NAS'12. IEEE, June 2012, pages 39–48.

[180] B. Menkus. Understanding the use of passwords. *Computers & Security*, 7(2), (1988), pages 132 – 136.

[181] N. Micallef, M. Just, L. Baillie, M. Halvey, and H. G. Kayacik. Why aren't users using protection? investigating the usability of smartphone locking. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '15, New York, NY, USA, (2015). ACM, pages 284–294.

[182] M. Mickisch. Designing a graphical shoulder surfing resistant authentication mechanism for off-the-shelf smartphones. Practical research course, Ludwig-Maximilians-Universität München, (2015).

[183] W. Moncur and G. Leplâtre. Pictures at the atm: exploring the usability of multiple graphical passwords. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, (2007), pages 887–894.

[184] R. Morris and K. Thompson. Password security: A case history. *Communications of the ACM*, 22(11), (1979), pages 594–597.

[185] I. Muslukhov, Y. Boshmaf, C. Kuo, J. Lester, and K. Beznosov. Understanding users' requirements for data protection in smartphones. In *28th International Conference on Data Engineering Workshops*, ICDEW'12. IEEE, (2012), pages 228–235.

[186] A. Mylonas, M. Theoharidou, and D. Gritzalis. Assessing privacy risks in android: A user-centric approach. In *Risk Assessment and Risk-Driven Testing*, volume 8418 of *Lecture Notes in Computer Science*. Springer International Publishing, (2014), pages 21–37.

[187] D. Nali and J. Thorpe. Analyzing user choice in graphical passwords. Technical report, School of Computer Science, Carleton University, (2004). TR-04-01.

[188] A. Narayanan and V. Shmatikov. Fast dictionary attacks on passwords using time-space tradeoff. In *Proceedings of the 12th ACM Conference on Computer and Communications Security*, CCS '05, New York, NY, USA, (2005). ACM, pages 364–372.

[189] D. L. Nelson, V. S. Reed, and J. R. Walling. Pictorial superiority effect. *Journal of Experimental Psychology: Human Learning and Memory*, 2(5), (1976), pages 523–528.

[190] N. H. Nie and L. Erbring. Internet and society: a preliminary report. In *The Digital Divide*. MIT Press, Cambridge, MA, USA, (2001), pages 269–271.

[191] R. Nithyanand and R. Johnson. The password allocation problem: Strategies for reusing passwords effectively. In *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society*, WPES '13, New York, NY, USA, (2013). ACM, pages 255–260.

[192] Y. Niu and H. Chen. Gesture authentication with touch input for mobile devices. In *Security and Privacy in Mobile Information and Communication Systems*. Springer, (2012), pages 13–24.

[193] D. A. Norman and S. W. Draper. User centered system design. *New Perspectives on Human-Computer Interaction, L. Erlbaum Associates Inc., Hillsdale, NJ*, (1986).

[194] G. Notoatmodjo and C. Thomborson. Passwords and perceptions. In *Proceedings of the Seventh Australasian Conference on Information Security - Volume 98*, AISC '09, Darlinghurst, Australia, (2009). Australian Computer Society, Inc., pages 71–78.

[195] I. Oakley and A. Bianchi. Multi-touch passwords for mobile device access. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, New York, NY, USA, (2012). ACM, pages 611–612.

[196] L. O'Gorman. Comparing passwords, tokens, and biometrics for user authentication. *Proceedings of the IEEE*, 91(12), Dec 2003, pages 2021–2040.

[197] P. C. v. Oorschot and J. Thorpe. On predictive models and user-drawn graphical passwords. *ACM Transactions on Information and System Security*, 10, TISSEC, January 2008, pages 5:1–5:33.

[198] A. Paivio and K. Csapo. Picture superiority in free recall: Imagery or dual coding? *Cognitive Psychology*, 5(2), (1973), pages 176 – 206.

[199] V. M. Patel, R. Chellappa, D. Chandra, and B. Barbello. Continuous user authentication on mobile devices: Recent progress and remaining challenges. *IEEE Signal Processing Magazine*, 33(4), July 2016, pages 49–61.

[200] J. Pi, P. De, and K. Mueller. Using gpus to crack android pattern-based passwords. In *International Conference on Parallel and Distributed Systems*, ICPADS'13, Dec 2013, pages 450–451.

[201] M. L. Polla, F. Martinelli, and D. Sgandurra. A survey on security for mobile devices. *IEEE Communications Surveys Tutorials*, 15(1), First 2013, pages 446–471.

[202] L. Y. Por, X. T. Lim, M. T. Su, and F. Kianoush. The design and implementation of background pass-go scheme towards security threats. *WSEAS Transactions on Information Science and Applications*, 5(6), June 2008, pages 943–952.

[203] R. Proctor, M.-C. Lien, K.-P. Vu, E. Schultz, and G. Salvendy. Improving computer security for authentication of users: Influence of proactive password restrictions. *Behavior Research Methods*, 34, 10.3758/BF03195438, (2002), pages 163–169.

[204] K. Renaud and A. De Angeli. Visual passwords: cure-all or snake-oil? *Communications of the ACM*, 52(12), Dec. 2009, pages 135–140.

[205] S. Riley. Password Security: What Users Know and What They Actually Do. *Usability News*, 8(1), Feb. 2006, pages 1–3.

[206] O. Riva, C. Qin, K. Strauss, and D. Lymberopoulos. Progressive authentication: Deciding when to authenticate on mobile phones. In *Proceedings of the 21st USENIX Security Symposium*, USENIX Security'12, Bellevue, WA, (2012). USENIX Association, pages 301–316.

[207] V. Roth, K. Richter, and R. Freidinger. A pin-entry method resilient against shoulder surfing. In *Proceedings of the 11th ACM conference on Computer and communications security*, CCS'04, New York, NY, USA, (2004). ACM, pages 236–245.

[208] N. Sae-Bae, K. Ahmed, K. Isbister, and N. Memon. Biometric-rich gestures: A novel approach to authentication on multi-touch devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, New York, NY, USA, (2012). ACM, pages 977–986.

[209] J. Saltzer and M. Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9), Sept 1975, pages 1278–1308.

[210] N. Sander, G. J. Abel, R. Bauer, and J. Schmidt. Visualising migration flow data with circular plots. *Vienna Institute of Demography Working Papers*, 2014-2, (2014), pages 1–35.

[211] H. Sasamoto, N. Christin, and E. Hayashi. Undercover: Authentication usable in front of prying eyes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, New York, NY, USA, (2008). ACM, pages 183–192.

[212] M. A. Sasse, S. Brostoff, and D. Weirich. Transforming the 'weakest link' a human/computer interaction approach to usable and effective security. *BT Technology Journal*, 19(3), July 2001, pages 122–131.

[213] K. Scarfone and M. Souppaya. Guide to enterprise password management. Technical report, National Institute of Standards and Technology, (2009). NIST Special Publication (SP) 800-118 (RETIRED DRAFT).

[214] F. Schaub, M. Walch, B. Könings, and M. Weber. Exploring the design space of graphical passwords on smartphones. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, New York, NY, USA, (2013). ACM, pages 11:1–11:14.

[215] S. Schechter, C. Herley, and M. Mitzenmacher. Popularity is everything: a new approach to protecting passwords from statistical-guessing attacks. In *Proceedings of the 5th USENIX conference on Hot topics in security*, HotSec'10, Berkeley, CA, USA, (2010). USENIX Association, pages 1–8.

[216] R. Schlöglhofer and J. Sametinger. Secure and usable authentication on mobile devices. In *Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia*, MoMM '12, New York, NY, USA, (2012). ACM, pages 257–262.

[217] S. Schneegass, Y. Oualil, and A. Bulling. Skullconduct: Biometric user identification on eyewear computers using bone conduction through the skull. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, New York, NY, USA, (2016). ACM, pages 1379–1384.

[218] S. Schneegass, F. Steimle, A. Bulling, F. Alt, and A. Schmidt. Smudgesafe: Geometric image transformations for smudge-resistant user authentication. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, New York, NY, USA, (2014). ACM, pages 775–786.

[219] B. Schneier. Real-world passwords. *Schneier on Security*, schneier.com, (2006).

[220] J. Seifert, A. De Luca, B. Conradi, and H. Hussmann. TreasurePhone: Context-sensitive user data protection on mobile phones. In *Pervasive Computing: 8th International Conference*, Pervasive 2010. Springer Berlin Heidelberg, (2010), pages 130–137.

[221] R. Shadmehr and T. Brashers-Krug. Functional stages in the formation of human long-term motor memory. *The Journal of Neuroscience*, 17(1), (1997), pages 409–419.

[222] M. Shahzad, A. X. Liu, and A. Samuel. Secure unlocking of mobile touch screen devices by simple gestures: You can see it but you can not do it. In *Proceedings of the 19th Annual International Conference on Mobile Computing & Networking*, MobiCom '13, New York, NY, USA, (2013). ACM, pages 39–50.

[223] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1), Jan. 2001, pages 3–55.

[224] R. Shay, S. Komanduri, A. L. Durity, P. S. Huh, M. L. Mazurek, S. M. Segreti, B. Ur, L. Bauer, N. Christin, and L. F. Cranor. Can long passwords be secure and usable? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, New York, NY, USA, (2014). ACM, pages 2927–2936.

[225] R. Shay, S. Komanduri, P. G. Kelley, P. G. Leon, M. L. Mazurek, L. Bauer, N. Christin, and L. F. Cranor. Encountering stronger password requirements: User attitudes and behaviors. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, New York, NY, USA, (2010). ACM, pages 2:1–2:20.

[226] R. N. Shepard. Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, 6(1), (1967), pages 156 – 163.

[227] M. Sherman, G. Clark, Y. Yang, S. Sugrim, A. Modig, J. Lindqvist, A. Oulasvirta, and T. Roos. User-generated free-form gestures for authentication: Security and memorability. In *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '14, New York, NY, USA, (2014). ACM, pages 176–189.

[228] B. Shneiderman. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 3rd edition, (1997).

[229] H. Siadati, P. Gupta, S. Smith, N. Memon, and M. Ahamad. Fortifying android patterns using persuasive security framework. In *The Ninth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, UBICOMM 2015. IARIA, (2015), pages 1–8.

[230] S. Singh, A. Cabraal, C. Demosthenous, G. Astbrink, and M. Furlong. Password sharing: Implications for security design based on social practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, New York, NY, USA, (2007). ACM, pages 895–904.

[231] L. Sjoeberg. Factors in risk perception. *Risk Analysis*, 20(1), (2000), pages 1–12.

[232] Y. Song, G. Cho, S. Oh, H. Kim, and J. H. Huh. On the effectiveness of pattern lock strength meters: Measuring the strength of real world pattern locks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, New York, NY, USA, (2015). ACM, pages 2343–2352.

[233] M. Souppaya and K. Scarfone. Guidelines for managing the security of mobile devices in the enterprise. Technical report, National Institute of Standards and Technology, June 2013. NIST Special Publication 800-124.

[234] F. Stajano. One user, many hats; and, sometimes, no hat: Towards a secure yet usable pda. In *Security Protocols*, volume 3957 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, (2006), pages 51–64.

[235] L. Standing. Learning 10,000 pictures. *The Quarterly journal of experimental psychology*, 25(2), May 1973, pages 207–222.

[236] L. Standing, J. Conezio, and R. Haber. Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic Science*, 19(2), (1970), pages 73–74.

[237] G. Stenberg, K. Radeborg, and L. Hedman. The picture superiority effect in a cross-modality recognition task. *Memory & Cognition*, 23(4), (1995), pages 425–441.

[238] E. Stobert and R. Biddle. The password life cycle: User behaviour in managing passwords. In *Proceedings of the 10th Symposium on Usable Privacy and Security*, Soups'14. USENIX Association (2014), (2014), pages 243–255.

[239] E. Stobert, A. Forget, S. Chiasson, P. C. van Oorschot, and R. Biddle. Exploring usability effects of increasing security in click-based graphical passwords. In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, New York, NY, USA, (2010). ACM, pages 79–88.

[240] W. C. Summers and E. Bosworth. Password policy: The good, the bad, and the ugly. In *Proceedings of the Winter International Synposium on Information and Communication Technologies*, WISICT '04. Trinity College Dublin, (2004), pages 1–6.

[241] C. Sun, Y. Wang, and J. Zheng. Dissecting pattern unlock: The effect of pattern strength meter on pattern selection. *Journal of Information Security and Applications*, 19(4–5), (2014), pages 308–320.

[242] J. Sun, R. Zhang, J. Zhang, and Y. Zhang. Touchin: Sightless two-factor authentication on multi-touch mobile devices. In *Conference on Communications and Network Security*, CNS'14. IEEE, Oct 2014, pages 436–444.

[243] X. Suo. A study of graphical password for mobile devices. In *Mobile Computing, Applications, and Services*, volume 130 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*. Springer International Publishing, (2014), pages 202–214.

[244] T. Takada. Fakepointer: An authentication scheme for improving security against peeping attacks using video cameras. In *Mobile Ubiquitous Computing, Systems, Services and Technologies, 2008. UBICOMM '08. The Second International Conference on*, Sept 2008, pages 395–400.

[245] T. Takada, T. Onuki, and H. Koike. Awase-e: Recognition-based image authentication scheme using users' personal photographs. In *Innovations in Information Technology, 2006*, Nov 2006, pages 1–5.

[246] D. S. Tan, P. Keyani, and M. Czerwinski. Spy-resistant keyboard: More secure password entry on public touch screen displays. In *Proceedings of the 17th Australia Conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future*, OZCHI '05, Narrabundah, Australia, (2005). Computer-Human Interaction Special Interest Group (CHISIG) of Australia, pages 1–10.

[247] V. Taneski, M. Hericko, and B. Brumen. Password security x2014; no change in 35 years? *Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2014, pages 1360–1365.

[248] H. Tao and C. Adams. Pass-go: A proposal to improve the usability of graphical passwords. *IJ Network Security*, 7(2), (2008), pages 273–292.

[249] F. Tari, A. A. Ozok, and S. H. Holden. A comparison of perceived and real shoulder-surfing risks between alphanumeric and graphical passwords. In *Proceedings of the Second Symposium on Usable Privacy and Security*, SOUPS '06, New York, NY, USA, (2006). ACM, pages 56–66.

[250] M. Theoharidou, A. Mylonas, and D. Gritzalis. A risk assessment method for smartphones. In *Information Security and Privacy Research*, volume 376 of *IFIP Advances in Information and Communication Technology*. Springer Berlin Heidelberg, (2012), pages 443–456.

[251] J. Thorpe, M. Al-Badawi, B. MacRae, and A. Salehi-Abari. The presentation effect on graphical passwords. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, New York, NY, USA, (2014). ACM, pages 2947–2950.

[252] J. Thorpe and P. Van Oorschot. Towards secure design choices for implementing graphical passwords. In *Computer Security Applications Conference, 2004. 20th Annual*. IEEE, (2004), pages 50–60.

[253] A. Treisman. Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing*, 31(2), (1985), pages 156 – 177.

[254] S. Uellenbeck, M. Dürmuth, C. Wolf, and T. Holz. Quantifying the security of graphical passwords: The case of android unlock patterns. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer &#38; Communications Security*, CCS '13, New York, NY, USA, (2013). ACM, pages 161–172.

[255] B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. L. Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, N. Christin, and L. F. Cranor. How does your password measure up? the effect of strength meters on password creation. In *Presented as part of the 21st USENIX Security Symposium*, USENIX Security 12, Bellevue, WA, (2012). USENIX, pages 65–80.

[256] T. Valentine. Memory for passfaces after a long delay. Technical report, Goldsmiths College, University of London, (1999).

[257] D. Van Bruggen, S. Liu, M. Kajzer, A. Striegel, C. R. Crowell, and J. D'Arcy. Modifying smartphone user locking behavior. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, New York, NY, USA, (2013). ACM, pages 10:1–10:14.

[258] W. van Eekelen, J. van den Elst, and V.-J. Khan. Dynamic layering graphical elements for graphical password schemes. In *Proceedings of the Chi Sparks 2014*

Conference: HCI Research, Innovation, and Implementation, Creating the Difference. The Hague University of Applied Sciences, (2014), pages 65–73.

[259] P. Van Oorschot, A. Salehi-Abari, and J. Thorpe. Purely automated attacks on passpoints-style graphical passwords. *IEEE Transactions on Information Forensics and Security*, 5(3), Sept 2010, pages 393–405.

[260] C. Varenhorst, M. Van Kleek, and L. Rudolph. Passdoodles: A lightweight authentication method. Technical report, Research Science Institute, Massachusetts Institute of Technology, (2004).

[261] M. Volkamer, K. Renaud, O. Kulyk, and S. Emeröz. A socio-technical investigation into smartphone security. In *Security and Trust Management*, volume 9331 of *Lecture Notes in Computer Science*. Springer International Publishing, (2015), pages 265–273.

[262] E. von Zezschwitz, A. De Luca, B. Brunkow, and H. Hussmann. Swipin: Fast and secure pin-entry on smartphones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, New York, NY, USA, (2015). ACM, pages 1403–1406.

[263] E. von Zezschwitz, A. De Luca, and H. Hussmann. Survival of the shortest: A retrospective analysis of influencing factors on password composition. In *Human-Computer Interaction – INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part III*. Springer Berlin Heidelberg, (2013), pages 460–467.

[264] E. von Zezschwitz, A. De Luca, and H. Hussmann. Honey, I shrunk the keys: Influences of mobile devices on password composition and authentication performance. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction*. ACM Press (2014), (2014), pages 461–470.

[265] E. von Zezschwitz, A. De Luca, P. Janssen, and H. Hussmann. Easy to draw, but hard to trace?: On the observability of grid-based (un)lock patterns. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, New York, NY, USA, (2015). ACM, pages 2339–2342.

[266] E. von Zezschwitz, P. Dunphy, and A. De Luca. Patterns in the wild: A field study of the usability of pattern and pin-based authentication on mobile devices. In *Proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services*, MobileHCI '13, New York, NY, USA, (2013). ACM, pages 261–270.

[267] E. von Zezschwitz, M. Eiband, D. Buschek, S. Oberhuber, A. De Luca, F. Alt, and H. Hussmann. On quantifying the effective password space of grid-based unlock gestures. In *Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia*, MUM '16, New York, NY, USA, (2016). ACM, pages 1–10.

[268] E. von Zezschwitz, A. Koslow, A. De Luca, and H. Hussmann. Making graphic-based authentication secure against smudge attacks. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, New York, NY, USA, (2013). ACM, pages 277–286.

[269] Y. Wang, K. Streff, and S. Raman. Smartphone security challenges. *Computer*, 45(12), (2012), pages 52–58.

[270] Z. Wang, J. Jing, and L. Li. Time evolving graphical password for securing mobile devices. In *Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security*, ASIA CCS '13, New York, NY, USA, (2013). ACM, pages 347–352.

[271] K. Watanabe, F. Higuchi, M. Inami, and T. Igarashi. Cursorcamouflage: Multiple dummy cursors as a defense against shoulder surfing. In *SIGGRAPH Asia 2012 Emerging Technologies*, SA '12, New York, NY, USA, (2012). ACM, pages 6:1–6:2.

[272] M. Weir, S. Aggarwal, M. Collins, and H. Stern. Testing metrics for password creation policies by attacking large sets of revealed passwords. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*, CCS '10, New York, NY, USA, (2010). ACM, pages 162–175.

[273] M. Weir, S. Aggarwal, B. de Medeiros, and B. Glodek. Password cracking using probabilistic context-free grammars. *Computer & Security*, IEEE (2009), (2009), pages 391–405.

[274] D. Weirich and M. A. Sasse. Pretty good persuasion: A first step towards effective password security in the real world. In *Proceedings of the 2001 Workshop on New Security Paradigms*, NSPW '01, New York, NY, USA, (2001). ACM, pages 137–143.

[275] R. Weiss and A. De Luca. Passshapes: Utilizing stroke based authentication to increase password memorability. In *Proceedings of the 5th Nordic Conference on Human-computer Interaction: Building Bridges*, NordiCHI '08, New York, NY, USA, (2008). ACM, pages 383–392.

[276] S. Wiedenbeck, J. Waters, J. Birget, A. Brodskiy, and N. Memon. Passpoints: Design and longitudinal evaluation of a graphical password system. *International Journal of Human-Computer Studies*, 63(1), (2005), pages 102–127.

[277] S. Wiedenbeck, J. Waters, J.-C. Birget, A. Brodskiy, and N. Memon. Authentication using graphical passwords: Effects of tolerance and image choice. In *Proceedings of the 2005 Symposium on Usable Privacy and Security*, SOUPS '05, New York, NY, USA, (2005). ACM, pages 1–12.

[278] S. Wiedenbeck, J. Waters, L. Sobrado, and J.-C. Birget. Design and evaluation of a shoulder-surfing resistant graphical password scheme. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '06, New York, NY, USA, (2006). ACM, pages 177–184.

[279] O. Wiese and V. Roth. Pitfalls of Shoulder Surfing Studies. In *Proceedings of the NDSS Workshop on Usable Security 2015*, USEC'15. Internet Society, (2015), pages 1–6.

[280] D. Wigdor, C. Forlines, P. Baudisch, J. Barnwell, and C. Shen. Lucid touch: A see-through mobile device. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*, UIST '07, New York, NY, USA, (2007). ACM, pages 269–278.

[281] C. Winkler, J. Gugenheimer, A. De Luca, G. Haas, P. Speidel, D. Dobbelstein, and E. Rukzio. Glass unlock: Enhancing security of smartphone unlocking through leveraging a private near-eye display. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, New York, NY, USA, (2015). ACM, pages 1407–1410.

[282] A. Wolin, B. Eoff, and T. Hammond. Shortstraw: A simple and effective corner finder for polylines. In *Proceedings of the Fifth Eurographics Conference on Sketch-Based Interfaces and Modeling*, SBM'08, Aire-la-Ville, Switzerland, Switzerland, (2008). Eurographics Association, pages 33–40.

[283] T.-S. Wu, M.-L. Lee, H.-Y. Lin, and C.-Y. Wang. Shoulder-surfing-proof graphical password authentication scheme. *International Journal of Information Security*, 13(3), (2014), pages 245–254.

[284] X. Xiao, T. Han, and J. Wang. Lensgesture: Augmenting mobile interactions with back-of-device finger gestures. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, New York, NY, USA, (2013). ACM, pages 287–294.

[285] T. Yamamoto, Y. Kojima, and M. Nishigaki. A shoulder-surfing-resistant image-based authentication system with temporal indirect image selection. In *Security and Management*, (2009), pages 188–194.

[286] J. Yan, A. Blackwell, R. Anderson, and A. Grant. Password memorability and security: Empirical results. *IEEE Security and Privacy*, 2(5), Sept. 2004, pages 25–31.

[287] J. J. Yan. A note on proactive password checking. In *Proceedings of the 2001 Workshop on New Security Paradigms*, NSPW '01, New York, NY, USA, (2001). ACM, pages 127–135.

[288] Q. Yan, J. Han, Y. Li, J. Zhou, and R. H. Deng. Designing leakage-resilient password entry on touchscreen mobile devices. In *Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security*, ASIA CCS '13, New York, NY, USA, (2013). ACM, pages 37–48.

[289] X.-D. Yang, E. Mak, P. Irani, and W. F. Bischof. Dual-surface input: Augmenting one-handed interaction with coordinated front and behind-the-screen input. In *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '09, New York, NY, USA, (2009). ACM, pages 5:1–5:10.

[290] Y. Yang, J. Lindqvist, and A. Oulasvirta. Text entry method affects password security. In *The LASER Workshop: Learning from Authoritative Security Experiment Results*, LASER 2014. USENIX Association, (2014), pages 11–20.

[291] H. Yoon, S. H. Park, and K. T. Lee. Exploiting ambient light sensor for authentication on wearable devices. In *Fourth International Conference on Cyber Security, Cyber Warfare, and Digital Forensic*, CyberSec'15, Oct 2015, pages 95–100.

[292] N. H. Zakaria, D. Griffiths, S. Brostoff, and J. Yan. Shoulder surfing defence for recall-based graphical passwords. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11, New York, NY, USA, (2011). ACM, pages 6:1–6:12.

[293] D. Zakay and R. A. Block. Prospective and retrospective duration judgments: An executive-control perspective. *Acta neurobiologiae experimentalis*, 64(3), (2004), pages 319–328.

[294] H. Zhao and X. Li. S3pas: A scalable shoulder-surfing resistant textual-graphical password authentication scheme. In *21st International Conference on Advanced Information Networking and Applications Workshops*, volume 2 of *AINAW '07*, May 2007, pages 467–472.

[295] Z. Zhao, G.-J. Ahn, and H. Hu. Picture gesture authentication: Empirical analysis, automated attacks, and scheme evaluation. *ACM Transactions on Information and System Security (TISSEC)*, 17(4), Apr. 2015, pages 14:1–14:37.

[296] B. B. Zhu, D. Wei, M. Yang, and J. Yan. Security implications of password discretization for click-based graphical passwords. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, Republic and Canton of Geneva, Switzerland, (2013). International World Wide Web Conferences Steering Committee, pages 1581–1591.

[297] J. Zimmerman, J. Forlizzi, and S. Evenson. Research through design as a method for interaction design research in hci. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, (2007), pages 493–502.

[298] M. Zviran and W. J. Haga. Password security: an empirical study. *Journal of Management Information Systems*, 15(4), Mar. 1999, pages 161–185.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig und ohne unerlaubte Beihilfe angefertigt wurde.

München, den 24. August 2016

Emanuel von Zezschwitz