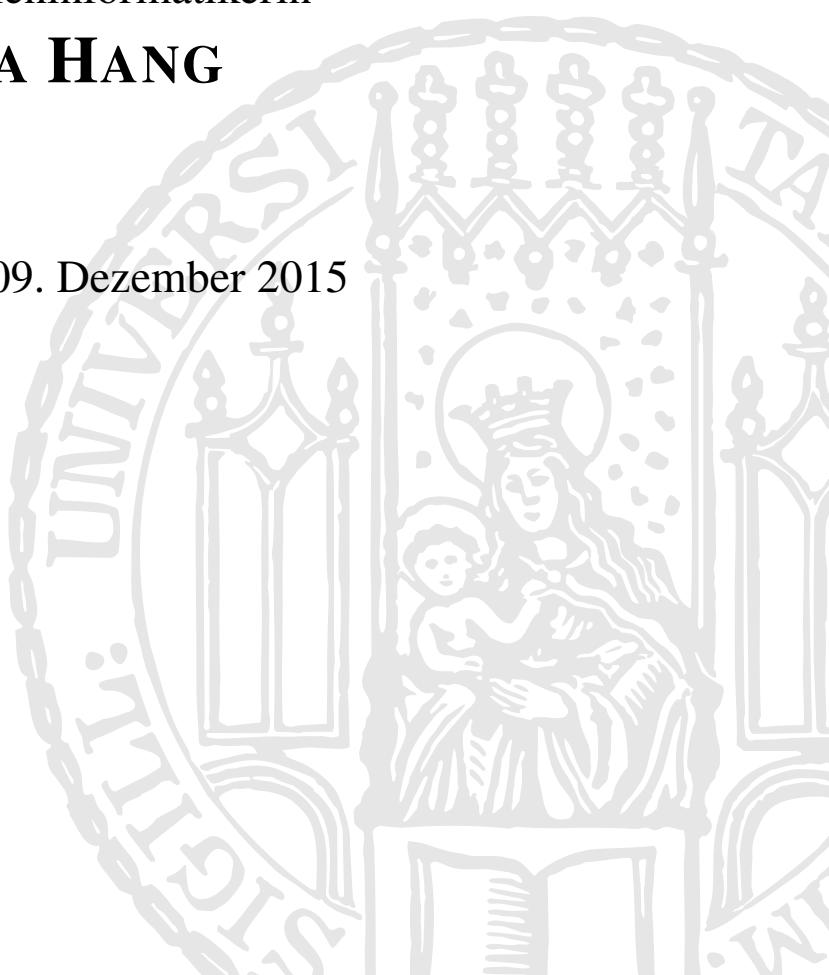

**EXPLOITING AUTOBIOGRAPHICAL MEMORY FOR
FALLBACK AUTHENTICATION ON SMARTPHONES**

DISSERTATION

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

vorgelegt von
Diplom-Medieninformatikerin
ALINA HANG

München, den 09. Dezember 2015



Erstgutachter: Prof. Dr. Heinrich Hussmann
Zweitgutachter: Associate Research Professor Lujo Bauer

Tag der mündlichen Prüfung: 3. März 2016

Abstract

Smartphones have advanced from simple communication devices to multipurpose devices that capture almost every single moment in our daily lives and thus contain sensitive data like photos or contact information. In order to protect this data, users can choose from a variety of authentication schemes. However, what happens if one of these schemes fails, for example, when users are not able to provide the correct password within a limited number of attempts? So far, situations like this have been neglected by the usable security and privacy community that mainly focuses on primary authentication schemes. But fallback authentication is comparably important to enable users to regain access to their devices (and data) in case of lockouts. In theory, any scheme for primary authentication on smartphones could also be used as fallback solution. In practice, fallback authentication happens less frequently and imposes different requirements and challenges on its design.

The aim of this work is to understand and address these challenges. We investigate the occurrences of fallback authentication on smartphones in real life in order to grasp the characteristics that fallback authentication conveys. We also get deeper insights into the difficulties that users have to cope with during lockout situations. In combination with the knowledge from previous research, these insights are valuable to provide a detailed definition of fallback authentication that has been missing so far. The definition covers usability and security characteristics and depicts the differences to primary authentication.

Furthermore, we explore the potential of autobiographical memory, a part of the human memory that relates to personal experiences of the past, for the design of alternative fallback schemes to overcome the well-known memorability issues of current solutions. We present the design and evaluation of two static approaches that are based on the memory of locations and special drawings. We also cover three dynamic approaches that relate to recent smartphone activities, icon arrangements and installed apps. This series of work allows us to analyze the suitability of different types of memories for fallback authentication. It also helps us to extend the definition of fallback authentication by identifying factors that influence the quality of fallback schemes.

The main contributions of this thesis can be summarized as follows: First, it gives essential insights into the relevance, frequency and problems of fallback authentication on smartphones in real life. Second, it provides a clear definition of fallback authentication to classify authentication schemes based on usability and security properties. Third, it shows example implementations and evaluations of static and dynamic fallback schemes that are based on different autobiographical memories. Finally, it discusses the advantages and disadvantages of these memories and gives recommendations for their design, evaluation and analysis in the context of fallback authentication.

Zusammenfassung

Aus vormals einfachen Kommunikationsgeräten haben sich Smartphones inzwischen zu Multifunktionsgeräten weiterentwickelt, die fast jeden einzelnen Moment in unserem Alltag verfolgen und aufzeichnen. So ist es nicht verwunderlich, dass diese Geräte auch viele sensible Daten beinhalten, wie zum Beispiel Fotos oder Kontaktinformationen. Um diese Daten zu schützen, können Smartphone-Nutzer aus einer Vielzahl von Authentifizierungsverfahren auswählen. Doch was passiert, wenn eines dieser Verfahren versagt, zum Beispiel wenn Nutzer nicht in der Lage sind ihr korrektes Passwort innerhalb einer begrenzten Anzahl von Versuchen einzugeben? Derartige Fragen wurden bislang von der *Usable Security und Privacy* Gemeinschaft vernachlässigt, deren Augenmerk vielmehr auf dem Forschungsfeld der primären Authentifizierung gerichtet war. Jedoch ist das Gebiet der Fallback-Authentifizierung von vergleichbarer Bedeutung, um Nutzern die Möglichkeit zu bieten, wieder Zugang zu ihren Daten und Geräten zu erlangen, wenn sie sich aussperren. Im Prinzip kann jedes primäre Authentifizierungsverfahren auch für die Fallback-Authentifizierung eingesetzt werden. Da letzteres in der Praxis jedoch viel seltener passiert, bringt der Entwurf neuer Verfahren für die Fallback-Authentifizierung neue Anforderungen und Herausforderungen mit sich.

Ziel dieser Arbeit ist es, diese Herausforderungen zu verstehen und herauszuarbeiten. Dazu haben wir untersucht, wie häufig sich Smartphone-Nutzer im Alltag aussperren, um darauf basierend die Hauptanforderungen für den Entwurf von Verfahren zur Fallback-Authentifizierung herzuleiten. Zudem konnten wir durch die Untersuchung ein tieferes Verständnis für die Probleme der Nutzer in solchen Situationen entwickeln. Zusammen mit den Erkenntnissen aus verwandten Arbeiten ermöglichen die Ergebnisse der Untersuchung eine detaillierte Definition für den Begriff der Fallback-Authentifizierung bereitzustellen und unter Berücksichtigung von Faktoren der Nutzerfreundlichkeit und Sicherheit deren Unterschiede zur primären Authentifizierung hervorzuheben.

Zudem haben wir die Möglichkeiten des autobiographischen Gedächtnisses für den Entwurf alternativer Verfahren zu Fallback-Authentifizierung exploriert. Das autobiographische Gedächtnis ist ein Teil des menschlichen Gehirns und besteht aus persönlichen Erinnerungen der Vergangenheit. Durch den persönlichen Bezug erscheinen diese Erinnerungen vielversprechend, um die Probleme bekannter Verfahren zu überwinden. Im Rahmen dieser Arbeit stellen wir deshalb zwei statische und drei dynamische Verfahren zur Fallback-Authentifizierung vor, die sich auf autobiographischen Erinnerungen stützen. Während sich die statischen Verfahren auf ortsbezogene Erinnerungen und das Anfertigen spezieller Zeichnungen konzentrieren, basieren die dynamischen Verfahren auf Erinnerungen der nahen Vergangenheit (z. B. Aktivitäten auf dem Smartphone, Anordnung von Anwendungen oder deren Installation). Die vorgestellten Konzepte erlauben nicht nur das Potential verschiedener autobiographischer Erinnerungen zu analysieren, sondern ermöglichen es auch Faktoren zu identifizieren, die einen Einfluss auf die Qualität der vorgestellten Konzepte haben und somit nützlich sind, um die Definition der Fallback-Authentifizierung zu erweitern.

Der wissenschaftliche Beitrag dieser Arbeit lässt sich wie folgt zusammenfassen: (1) Die Arbeit gibt einen wichtigen Einblick in die Relevanz, Häufigkeit und Probleme der Fallback-Authentifizierung im Alltag der Nutzer. (2) Sie stellt eine klare Definition für den Begriff der Fallback-Authentifizierung bereit, um Authentifizierungssysteme anhand verschiedener Eigenschaften wie Nutzerfreundlichkeit und Sicherheit zu klassifizieren. (3) Sie diskutiert die Vor- und Nachteile verschiedener autobiographischer Erinnerungen anhand von Beispieleimplementierungen und gibt darauf basierend Empfehlungen zu deren Nutzung und Evaluierung im Kontext der Fallback-Authentifizierung.

Disclaimer

This thesis is the result of several projects that I have accomplished in the past four years of my time as a PhD student. None of these projects would have been possible without the help of my wonderful colleagues and diligent students. To appreciate this collaboration, I decided to use the scientific plural in this thesis. A clear overview of my personal contribution to each of the projects will be given next.

Chapter 3: A Survey on Fallback Authentication

The content of this chapter is based on a bachelor thesis by Victoria Müller [127]. The idea for the project was developed by me. I defined the project's goals and set the frame for the bachelor thesis. The work was conducted in a collaborative manner and each step in the project was jointly discussed in weekly meetings. However, as a supervisor, I was the key decision maker on how to proceed in the project (e.g., survey structure, question design, study execution and study evaluation).

Part of this work was published at the *International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI 2015)* with co-authors Alexander De Luca, Emanuel von Zezschwitz, Manuel Demmler and Heinrich Hussmann [77]. The main corpus of the paper was written by me (except for the limitations section that was written by Emanuel von Zezschwitz). Revisions to the paper were also made by me and were based on the feedback by the co-authors.

Chapter 5: Fallback Authentication Based on Static Enrollments

The content of this chapter is based on two bachelor theses by Lara Hirschbeck and Michael Richter [89, 145].

The idea for the first project was developed by me. I defined the project's goals and set the frame for the bachelor thesis. The work was conducted in a collaborative manner and each step in the project was jointly discussed in weekly meetings, but I was the key decision maker on how to proceed in the project (e.g., prototype features, study design and data analysis).

The second project was jointly supervised by Alexander De Luca and me. Again, the work was conducted in a collaborative manner and each step in the project was jointly discussed in weekly meetings. Alexander De Luca and I were the key decision makers on how to proceed in the project (e.g., prototype features, study design and data analysis).

The results from this work were published at the *Symposium on Usable Privacy and Security (SOUPS 2015)* with co-authors Alexander De Luca, Michael Richter, Matthew Smith and Heinrich Hussmann [76]. The main corpus of the paper was written by me. Revisions to the paper were also made by me and were based on the feedback by the co-authors.

Chapter 6: Fallback Authentication Based on Implicit Enrollments

The content of this chapter is based on three main projects that, in turn, consist of several bachelor theses.

The first project on dynamic security questions was based on two bachelor theses by Stephan Thalhammer and Philipp Hauptmann [81, 173]. The idea for dynamic security questions originated from Emanuel von Zezschwitz and Alexander De Luca. This idea was then adapted to the context of fallback authentication by Alexander De Luca and me. The work was conducted in a collaborative manner with the corresponding students and each step in the project was jointly discussed in weekly meetings. Alexander De Luca and I were the key decision makers on how to proceed in the projects (e.g., prototype features, study design and data analysis).

Part of this project was published at the *International Conference on Human Computer Interaction (CHI 2015)* with co-authors Alexander De Luca and Heinrich Hussmann [75]. The main corpus of the paper was written by me and was constantly improved based on the feedback of the co-authors.

The second project on icon arrangements was based on two bachelor theses by John-Louis Gao and Carina Saliger [62, 150]. The idea for both theses was developed by me. The work was conducted in a collaborative manner with the students and each step in the project was jointly discussed in weekly meetings. However, I was the key decision maker on how to proceed in the project (e.g., prototype features, study design and data analysis).

Part of this project was published at the *International Conference on Human Computer interaction (CHI 2014)* with co-authors Alexander De Luca and Heinrich Hussmann [74]. The main corpus of the paper was written by me. Revisions to the paper were also made by me and were based on suggestions by the co-authors.

The last project on app installations was based on a bachelor thesis by Manuel Demmler [45]. The idea for the project was developed by me. I defined the project's goals and set the frame for the resulting bachelor thesis. The work was conducted in a collaborative manner and each step in the project was jointly discussed in weekly meetings, but I was the key decision maker on how to proceed in the project (e.g., prototype features, study design and data analysis).

The project was part of a paper that was published at the *International Conference on Human-Computer Interaction with Mobile Devices and Service (MobileHCI 2015)* with co-authors Alexander De Luca, Emanuel von Zezschwitz, Manuel Demmler and Heinrich Hussmann [77]. The main corpus of the paper was written by me (except for the limitations section that was written by Emanuel von Zezschwitz). Revisions to the paper were also made by me and were based on the feedback by the co-authors.

ACKNOWLEDGMENTS

The completion of this thesis was a long and emotional journey, filled with ups and downs. I am happy that I was surrounded by wonderful people who comforted me in moments of anxiety and with whom I could share moments of joy. All of them deserve my greatest thanks as I would not have been able to succeed without their invaluable support.

First and foremost, I would like to express my deepest gratitude to my supervisors: Thank you **Prof. Dr. Heinrich Hussmann** for your guidance, patience and constant support. You gave me the opportunity to pursue my research interests and provided me with excellent feedback whenever I needed it. Thank you **Associate Professor Lujo Bauer** for your genuine interest in my work. Our discussions and your feedback were a great encouragement and input to enhance the quality of this thesis.

I would also like to thank my wonderful colleagues, with whom I have shared so many memorable moments. **Alexander De Luca**: You have always been there for me throughout my time as a PhD student. You encouraged me to pursue ideas, prevented me from following stupid ones and helped me wherever you could. I am happy to have you as a friend. **Enrico Rukzio**: You introduced me to the world of research which I entered with fascination. The first publication, the first conference and the first talk were all the results of my unforgettable time in Lancaster. Thank you for making all of this possible. **Sebastian Löhmann**: You were not only my personal concert bodyguard, but you were also the sweetest office mate I could have ever imagined. Thank you for your small gifts that came so unexpectedly to brighten my day! **Simon Stusak**: This may sound weird, but you are the meanest nice guy. Although you were teasing me all the time, I am grateful for your soothing words when I needed them the most. Good news: your karma is saved. **Sarah Tausch**: You are the twin I never had. Together, we share the best fashion sense, the best yoga moves and the best office aliases. I hope we will never loose this special bond. **Emanuel von Zezschwitz**: You are the best example that there can be a balance between work and fun. Thank you for turning every event into a great party. You made unpacking an iPhone even more exciting than it already was. **Fabian Hennecke**: Your witty humor, your love for dinosaurs and your passion for funny links have often made my day. Please do not stop sending me cute cat pictures. **Sara Hennecke**: Thank you for teaching me that persistence is key for pursuing a PhD degree. Your words have often popped up my mind and they were proven to be true. **Doris Hausen**: You were the first to welcome me on my first day at the office. You introduced me to the Barkeeper, the Spamchannel and kept me posted about the latest gossip. Thank you for being my roommate all over the world. **Max-Emanuel Maurer**: I was (and still am) impressed by your skill to complete so many things, so fast, in so little time. However, I am glad that you were not as fast with your idea to ignite the burning paste. **Julie Wagner**: Thank you for sharing your tent with me at the Southside festival. It was a unique experience and I think that we cannot get any closer than that. **Hendrik Richter and Bettina Conradi**: Thank you for the brainstorming sessions in which we tried to find my research direction. Although I took a different path in the end, your input helped me a lot. **Axel Hösl**: You taught me the art of high five, showed me how to be a ninja and, most importantly, created my second

personality: Adrenalina. You will always be my Spaxl. **Tobias Seitz**: You are the best music teacher that one can have! Thank you for the wonderful guitar lessons. **Gregor Broll** and **Alexander Wiethoff**: It was a great pleasure to have you as supervisors for my diploma thesis. I appreciate the freedom that you gave me to be creative and to explore different ideas. **Sebastian Boring** and **Dominikus Baur**: I will always remember you as Statler and Waldorf. It was nice to have you as my roomies. **Raphael Wimmer**: For me, you represent the perfect image of a researcher. You are full of ideas, curious and have the drive to create something new. Thank you for sharing some of your ideas with me for my research on projector phones. **Henri Palleis**: I am fascinated by your calmness, attentiveness and dry sense of humor. Please keep these good traits of yours. **Daniel Buschek**: I would not have been surprised, if you had finished your PhD before me, but thank you for letting me finish first. As the time went by and I grew older, I saw senior PhD students leaving with their degrees and new ones coming. **Maria Fysaraki**, **Mohamed Khamis**, **Hanna Schneider** and all other "newbies": I wish you all the best for your research.

Furthermore, I would like to thank **Prof. Dr. Andreas Butz**, **Prof. Dr. Michael Rohs** and **Prof. Dr. Florian Alt** for their feedback on my work during various occasions, such as the IDC in Venice. I would also like to give special thanks to the good souls of our research group who ensure a smooth operation of our daily business. **Franziska Schwamb**: I highly value your straightforwardness and sincerely thank you for always having time for me when I needed your help. **Rainer Fink**: Thank you for feeding my shopping addiction with all the vouchers.

In addition, I would like to acknowledge all external PhD students that visited our group once in a while. Without you, I would not have learned how the car of the future will look like or how we can use our body to perform text input.

These acknowledgments would not be complete if I did not mention the talented students that I have met and supervised in the course of my work. I appreciate their efforts to support me in my research. For that, I would like to express my special thanks to Michael Richter, Stephan Thalhammer, Philipp Hauptmann, John-Louis Gao, Carina Saliger, Manuel Demmler, Lara Hirschbeck and Victoria Müller.

Last, but certainly not least, I would like to thank my parents who did not spare any effort to support me in any of my endeavors. They gave me all the possibilities that they did not have. Without them, I would not be in the place I am today. Therefore, I dedicate this thesis to my parents with a few accompanying words in Vietnamese:

Lời cảm tạ quan trọng nhất con xin gửi đến ba mẹ: Hai người không chỉ ủng hộ con trong quãng thời gian gần đây, mà hơn nữa, đã luôn khuyến khích con trong tất cả mọi việc từ lúc con còn nhỏ cho đến bây giờ. Vì vậy, bài luận văn này con xin gửi tặng đến hai người để biểu lộ sự kính trọng của con đối với ba mẹ.

Table of Contents

List of Figures	xvii
List of Tables	xix
1 When One Door Closes, Another Opens	1
1.1 Losing the Key	2
1.1.1 Lockout Scenarios	2
1.1.2 Current Solutions	3
1.2 Problem Statement	4
1.3 Main Contributions	5
1.3.1 Insights into Mobile Fallback Authentication	6
1.3.2 Framework for Fallback Authentication	6
1.3.3 Analysis of Autobiographical Memories	6
1.3.4 Design and Evaluation of Fallback Schemes	7
1.4 Thesis Overview	7
2 Learning from the Past and Present	9
2.1 Passwords, the Good and Evil	10
2.1.1 Password Advice	10
2.1.2 Password Policies	10
2.1.3 Password Meters	11
2.1.4 Short Summary	11
2.2 Current Practices in Fallback Authentication	12
2.2.1 Fallback Authentication between 2007 and 2009	12
2.2.2 Fallback Authentication in 2014	13
2.2.3 Password Managers	13
2.2.4 Short Summary	14
2.3 The Design of Security Questions	14
2.3.1 Classification	14
2.3.2 The Usability of Security Questions	15

2.3.3	The Security of Security Questions	17
2.3.4	Advancements in Fallback Solutions	18
2.3.5	Short Summary	20
2.4	Smartphone Authentication	21
2.4.1	Passwords and PINs	21
2.4.2	Graphical Passwords	21
2.4.3	Biometric Passwords	22
2.4.4	Short Summary	22
2.5	Fallback Authentication on Smartphones	23
2.6	The Human Memory	23
2.6.1	The Multi-Store Model of Memory	23
2.6.2	Long-Term Memory	24
2.6.3	Autobiographical Memory	25
2.6.4	Short Summary	26
2.7	Lessons Learned	26
3	A Field Study on Mobile Fallback Authentication	29
3.1	Fallback Authentication on Smartphones	30
3.1.1	Cellular Network Access	30
3.1.2	Device Unlock	30
3.1.3	Account Login	31
3.2	Research Strategy	31
3.3	General Overview of Fallback Experiences	32
3.3.1	Survey Design	32
3.3.2	Coding	32
3.3.3	Participants	33
3.3.4	Results	34
3.4	Individual Reports on Fallback Experiences	35
3.4.1	Interview Design	36
3.4.2	Participants	36
3.4.3	Results	36
3.5	Discussion	38
3.5.1	Lack of Information	38
3.5.2	Feeling Secure, but Annoyed	39
3.5.3	Third-Party Dependencies	39
3.5.4	Third Parties	39
3.6	Lessons Learned	40

TABLE OF CONTENTS

4 A Framework for Fallback Authentication	41
4.1 Fallback Authentication	42
4.1.1 The Authentication Chain	42
4.1.2 Classification of Fallback Authentication	42
4.1.3 Definition of Fallback Authentication	45
4.2 Exploring the Design Space	46
4.2.1 Type of Replacement	46
4.2.2 Type of Enrollment	46
4.2.3 Type of Challenge and Type of Response	46
4.2.4 Overview of Example Implementations	46
4.3 Research Strategy	48
4.3.1 Usability Evaluation Methodology	48
4.3.2 Security Evaluation Methodology	49
4.3.3 Accuracy Metric	50
4.4 Chapter Summary	51
5 Fallback Authentication Based on Static Enrollments	53
5.1 Sketch-based Fallback Authentication	54
5.1.1 Approach	54
5.1.2 Design of Predefined Sketches	55
5.1.3 Prototype	55
5.1.4 Threat Model	56
5.1.5 User Study	57
5.1.6 Discussion	61
5.2 Location-based Questions	62
5.2.1 Approach	62
5.2.2 Question Design	63
5.2.3 Prototype	65
5.2.4 Threat Model	66
5.2.5 User Study	66
5.2.6 Discussion	75
5.3 Lessons Learned	77
6 Fallback Authentication Based on Dynamic Enrollments	79
6.1 Smartphone Activities	80
6.1.1 Question Design	80
6.1.2 Prototype	81
6.1.3 Threat Model	82

6.1.4	User Study	82
6.1.5	Follow-Up Study	86
6.1.6	Discussion	90
6.2	Icon Arrangements	92
6.2.1	Approach	92
6.2.2	Prototype	94
6.2.3	Threat Model	94
6.2.4	User Study	95
6.2.5	Follow-Up Study	100
6.2.6	Discussion	105
6.3	Installed Apps	107
6.3.1	Approach	108
6.3.2	Prototype	108
6.3.3	Threat Model	109
6.3.4	User Study	110
6.3.5	Discussion	115
6.4	Lessons Learned	116
7	Assembling the Pieces	119
7.1	Motivation for Fallback Authentication	120
7.2	Designing for Fallback Authentication	120
7.2.1	Type of Authentication Scheme	120
7.2.2	Autobiographical Memories	121
7.2.3	Type of Enrollment	122
7.2.4	Type of Challenge, Type of Response	123
7.3	Evaluating Fallback Authentication	124
7.3.1	Usability Evaluation	125
7.3.2	Security Evaluation	126
7.3.3	Other Evaluation Criteria	126
7.3.4	Order of Importance	127
7.3.5	Analyzing Fallback Authentication	127
7.4	Application Areas	128
7.5	Lessons Learned	128
8	Looking Back, Moving Forward	131
8.1	Contribution Summary	132
8.1.1	Understanding Fallback Authentication	132

TABLE OF CONTENTS

8.1.2	Definition of Fallback Authentication	132
8.1.3	Autobiographical Memories	134
8.1.4	Design of Fallback Schemes	134
8.1.5	Evaluation of Fallback Schemes	135
8.2	Limitations and Future Work	135
8.2.1	In-Depth Insights on Mobile Fallback Authentication	135
8.2.2	Different User Types	136
8.2.3	Development Platform	136
8.2.4	Beyond Episodic Memories	136
8.2.5	Real-World Deployment	137
8.3	Closing Remarks	137
A	Appendices	139
A.1	Overview of Fallback Schemes in 2014	140
Bibliography		143

List of Figures

2.1	Overview of different types of support for password creation	11
2.2	Overview of fallback schemes used on Alexa's top websites in 2014	13
2.3	Structure of human memory by Atkinson and Shiffrin	24
2.4	Classification of human memory and the construction of autobiographical memory	25
3.1	Example process of an authentication deadlock	31
4.1	The fallback authentication chain	43
5.1	Screenshots of the prototype for the study on 2-finger sketches	56
5.2	Screenshots of the prototype for the study on location-based security questions	65
5.3	Answer distributions during enrollment for location-based security questions	70
6.1	Screenshots of the prototype for the study on activity-based security questions	83
6.2	Overview of the number of correct answers for the first study on activity-based security questions	84
6.3	Overview of the number of correct answers for the follow-up study on activity-based security questions	88
6.4	Screenshots of the prototype for the study on icon arrangements on smartphones	93
6.5	Confidence ratings by users and adversaries about their submitted answers for the study on icon arrangements	99
6.6	Confidence ratings by users and adversaries about their submitted answers for the follow-up study on icon arrangement	104
6.7	Screenshots of the prototype for the study on installed apps	109
6.8	Guessability ratings by users for different types of adversaries for the study on installed apps	113

List of Tables

2.1	Overview of example security questions used by AOL and Yahoo in 2014	15
2.2	Overview of recall rates and guessing rates of security questions from different research	17
3.1	Demographic information about 244 online survey participants	33
3.2	Overview of the number of lockout experiences (online survey)	34
3.3	Overview of the number of lockout experiences (semi-structured interviews)	37
4.1	Design space for fallback authentication	47
5.1	Overview of nine 2-finger sketches	55
5.2	Parameter combinations for the analysis of the 2-finger sketches	58
5.3	Accuracy results for the 2-finger sketches	59
5.4	List of location-based security questions	64
5.5	Number of attempts and number of correct answers for location-based security questions	71
5.6	Accuracy results for location-based questions	74
6.1	List of activity-based security questions (first study)	81
6.2	Number of correct answers per question category (first study)	85
6.3	List of activity-based security questions (follow-up study)	87
6.4	Number of correct answers per question category (follow-up study)	89
6.5	Number of correct answers for the icon arrangements study	98
6.6	Number of correct answers for the icon arrangements study (follow-up)	103
6.7	Accuracy results for the icon arrangements study (follow-up)	105
6.8	Number of correctly identified apps	111
6.9	Accuracy results for the study on installed apps	115
7.1	Comparison of results	121

1

When One Door Closes, Another Opens

Imagine a girl who joyfully descends from the airplane. It is her first travel to Denmark and she is excited. She and her friend are now on the way to the train that is supposed to bring them into the city. They chatter in anticipation, making plans for their stay. As the train has not yet arrived, the girl wants to quickly drop a message to her parents that the airplane has landed safe and sound. She picks up her phone and turns it on as she had switched it off for the flight. Still focused on her friend's chatter, she notices that the phone requires some kind of PIN, but does not pay much attention to it. She repeatedly enters her lock screen PIN despite the notifications that pop up to inform her that the input was wrong. Shifting her attention back from her friend to the phone, she realizes that she just locked her SIM card. She had mixed up the PINs. What is she supposed to do?

1.1 Losing the Key

Losing access to our smartphone or parts of it is comparable to losing the key to our homes and thus to personal items like photos or financial records (e.g., [78, 105, 128]). While smartphones were mainly used for communication purposes in their early stages, they have transformed to multipurpose devices, supporting users in different tasks, ranging from scheduling appointments, initiating financial transactions to simply entertaining the user (e.g., [13, 151]). In order to protect the data that comes with these kinds of tasks, users can choose from a variety of authentication schemes like PINs or passwords. However, these authentication schemes as well as their users are not infallible and make the availability of alternative schemes as fallback options indispensable.

These alternatives are described by different terminology in the literature, such as last resort authentication (e.g., [155]), backup authentication (e.g., [156]), password reset (e.g., [60]) or fallback authentication (e.g., [139]). In the remainder of this thesis, we will use the term fallback authentication for several reasons: First, we do not consider the use of alternative authentication options to be a last resort solution. In fact, the last resort in a mobile context is resetting the phone to factory settings that, in turn, is coupled with the deletion of all data on the device. Second, the term backup authentication is ambiguous as it also refers to the authentication on backup systems that has nothing to do with lockout experiences. Third, although the term password reset describes the use of alternative authentication schemes, we were striving for a more general term that does not have a strong association with password resets for online accounts.

The remainder of this chapter is structured as follows: It introduces different scenarios in which lockouts may occur in a mobile context (Section 1.1.1). It further gives an overview of current fallback solutions to establish a basic understanding of fallback authentication in practice (Section 1.1.2). This forms the basis to introduce our problem statement (Section 1.2) and main contributions (Sections 1.3 and 1.4).

1.1.1 Lockout Scenarios

Lockouts refer to situations in which the user is no longer able to authenticate via the primary authentication scheme. They can, for example, happen when users are not able to recall their passwords or when they repeatedly enter the incorrect passwords until the number of allowed authentication attempts is exhausted and the primary scheme gets blocked. In a mobile context, there exist different situations that can lead to lockouts.

SIM Card Lockouts

The introducing scenario of this chapter was an example for a SIM card lockout. SIM cards are used to authenticate smartphones with the cellular network which, in turn, is required to enable telephony-related functionality (e.g., making calls). Since these cards store sensitive

data about the mobile subscriber, they can be protected by a PIN which, however, has to be entered in special circumstances only, for example, when the device is restarted. Due to the infrequent input, these situations are likely to cause potential lockouts due to forgotten PINs.

Lock Screen Lockouts

Lockouts can also occur during lock screen authentication on smartphones due to different circumstances. For example, biometric schemes are prone to false negatives that are not necessarily caused by the user, but are due to external factors, such as bad lighting [41]. In these situations, users depend on alternative means of authentication to regain access. Lockouts can also happen with knowledge-based schemes, such as PINs or graphical passwords, for example, when users enter them incorrectly multiple times in a row until the number of allowed attempts is exhausted.

Account Lockouts

Most smartphones enable users to access their online accounts on the go, either using a mobile browser or dedicated apps for these services (e.g., email or social networks). Due to the inconvenience of text input on smartphones, users often have the possibility to store their passwords for automated login. However, situations like spontaneous device sharing sometimes require users to log out of their accounts (e.g., when a friend borrows the phone to use an app [78]). This may lead to potential lockouts due to the lack of input training.

1.1.2 Current Solutions

For each of the described lockout scenarios, there are different fallback solutions. In this section, we will focus on solutions that are actually used in practice. This is done to provide a basic overview of how fallback authentication works and what problems they may cause. However, the main focus of this thesis aims at developing a profound understanding of these problems from a scientific point of view. This has been rarely done in the research community and we will elaborate on this later in more detail (see Section 1.2).

Solutions for SIM Card Lockouts

Personal unblocking codes (PUC) are often used for SIM card lockouts [148] and consist of up to eight digits. They are usually sent to the user by postal mail or can be looked up online. As the last resort (i.e., when the PUC is lost), users have to ask their service provider for replacement.

Solutions for Lock Screen Lockouts

With respect to lock screen lockouts, the solutions vary and depend on the operating system of the device as well as the type of authentication scheme used. For example, Android users

with Face Unlock¹ as their primary scheme are required to set up a graphical pattern as an alternative in case of failure. If this pattern, in turn, is entered incorrectly multiple times in a row, users have to wait for a certain amount of time before re-authentication with the same scheme is possible. However, Android users that have their device connected to an email account can circumvent these waiting times by performing email-based resets. As the last resort, users can reset their devices to factory settings. However, this may be coupled with the loss of their data if no backup was made beforehand.

In turn, iPhone users can set up a fingerprint-based authentication method called Touch ID.² In case Touch ID is activated, users have to provide an alternative PIN for fallback authentication. If this PIN is repeatedly entered incorrectly, the system introduces a delay before another authentication attempt is possible. To circumvent these waiting times, users can connect their smartphone to Apple's iTunes software. However, this method is coupled with data loss if no backup has been made.

There are some operating systems that do not offer any fallback options at all so that resetting the device is the only possibility. This is, for example, of concern for Windows Phone users or Blackberry users who do not have business accounts.

Solutions for Account Lockouts

With respect to account lockouts, the alternatives are often adopted from the desktop environment and include, for example, email-based password resets or the use of security questions (see Chapter 2).

1.2 Problem Statement

While it is comforting to know that solutions exist for most lockout situations, it is, at the same time, unsatisfying that so little is known about how and when they are used. Since mobile devices are often used on the go, current solutions may not work in all circumstances. For example, the retrieval of tokens or the use of email-based resets is difficult when users are not at home or Internet access is limited. In addition to this, some required tools, such as computers, cables or software, may not be at hand to perform fallback authentication.

So far, these potential problems have received little attention by the research community: there is no scientific data available that assesses the reasons that lead to lockout situations and how users cope with them. Hence, it is not possible to decide what users need from fallback solutions to recover from lockout situations and how these solutions should look like. Instead, research in the field of usable security and privacy mostly focuses on primary authentication. Fallback authentication (in particular on mobile devices) is often mentioned

¹ <https://support.google.com/nexus/answer/2781894?hl=en> (*last accessed 23/11/2015*)

² <https://support.apple.com/en-us/HT201371> (*last accessed 23/12/2015*)

as a side note only, for example, when proposed schemes do not work for primary authentication due to long authentication times (e.g., [83]).

The research that is available for fallback authentication is limited to the analysis and design of security questions (see Chapter 2). The key insights from this research is that security questions have very bad usability and security properties: questions that are easy to remember are often easy to guess, while questions that are hard to guess are often hard to remember (e.g., [68, 101, 102, 139, 139]). Therefore, it is discouraged to use them, as is, for fallback authentication. Although some endeavors have been made to design alternatives, finding appropriate solutions remains a challenging task.

One of the main shortcomings of security questions is that most of them rely on personal facts, such as facts about family members (e.g., mother's maiden name), pets (e.g., name of first cat) or personal preferences (e.g., favorite actor). Although easy to remember, these kinds of information are also often known by persons close to the user or even researchable by strangers through the help of public records or social networks [68, 101, 139].

However, what has been overlooked so far for the design of security questions is that autobiographical information goes beyond personal facts. On the contrary, autobiographical memory is a collection of memories from different brain structures, such as the semantic memory, episodic memory or memory for procedural skills and habits. These memories are all centered around the self and remembered more vividly than just facts [51, 176]. Episodic memories, for example, are not learned by heart, but instead are experienced through personal events and recalled more vividly [175]. Another example are procedural skills. They are often acquired through training and also easier to recall when they become an automated routine [32]. All things considered, autobiographical memories seem to be a great fit for the design of fallback schemes. Therefore, we analyze in this thesis how well autobiographical memories work in terms of usability (i.e., memorability) and security for fallback authentication to support users in lockout situations.

In summary, there are three major problems that we want to address. The first problem is the lack of understanding of why lockout situations occur and how well users cope with these situations. The second problem refers to the missing design space for fallback authentication: It is unclear which requirements should be met for the design of fallback schemes, for example, to distinguish them from primary authentication schemes. The third problem is related to the narrow view on autobiographical memory: Most security questions focus on personal facts so that other types of autobiographical memories have been seldom explored for fallback authentication.

1.3 Main Contributions

As previously discussed, the topic of fallback authentication has only been marginally touched by previous research and there exists, in particular for the mobile context, little knowledge about how well fallback solutions work, what kind of problems users have to

cope with and how alternative fallback schemes should look like. The main objective of this thesis is to provide a fundamental understanding on this matter. In order to achieve this, we address several research questions on our way towards this goal.

1.3.1 Insights into Mobile Fallback Authentication

While there exist some numbers on the lockout frequency of online accounts (e.g., [166]), there is no documentation available from manufacturers, service providers or research about how often fallback authentication is required on smartphones. This thesis provides some of the first solidly researched answers on this matter by using established research tools (i.e., surveys and semi-structured interviews) to analyze the frequency with which users experience different lockout situations on their smartphones (i.e., SIM card lockouts, lock screen lockouts and account lockouts). In addition to this, we accompany the raw numbers by personal anecdotes that allow to understand why certain problems arise and why certain countermeasures are taken by users during fallback authentication. Overall, these insights highlight the basic requirements for the design of mobile fallback schemes (see Chapter 3).

1.3.2 Framework for Fallback Authentication

We further provide a framework for fallback authentication to support usable security and privacy researchers and practitioners in classifying their authentication schemes (Chapter 4). The framework encompasses a clear definition of fallback authentication to depict the main differences to primary authentication. It further analyzes the requirements for the design and evaluation of fallback schemes. These requirements are derived from established usability and security criteria and are later extended by the results presented in this work. In addition to this, we suggest the accuracy measure as a metric to find the best trade-off between memorability and usability for knowledge-based authentication schemes. This metric uses the ratio of the number of correct decisions made by a system to the number of all decisions made to indicate how well a system (in our case a fallback scheme) works. We provide reasons in favor and against the use of this measure in the context of knowledge-based schemes and set an example by using it for the evaluation of our implementations.

1.3.3 Analysis of Autobiographical Memories

Since the human memory has more to offer than just personal facts, we review existing work in this domain to identify other types of autobiographical memories for the design of fallback schemes to overcome the shortcomings of traditional security questions (see Chapter 2). Through the design and implementation of five example prototypes, this thesis analyzes the potential and challenges of different memory types that include the skill to draw, episodic memories with a strong spatio-temporal context as well as episodic memories about recent

smartphone activities, icon arrangements and installed apps. In particular, this thesis shows how well/bad certain types of memories can be recalled by users and how secure/insecure certain types of memories are against attacks by different adversaries. In addition to this, we identify the memory types that have the best trade-off between the two factors, as the best memories in terms of memorability are not necessarily the best in terms of security.

1.3.4 Design and Evaluation of Fallback Schemes

In addition to the evaluation of autobiographical memories, we further show how they can be used for the implementation and evaluation of an actual fallback scheme. In particular, we explore different design options, such as the type of enrollment used. While static enrollments require users to provide the needed information in advance (i.e., before fallback authentication takes place), dynamic enrollments collect the needed information implicitly in the background. Both approaches have their advantages and disadvantages that are discussed in this thesis based on our study results. We further demonstrate how authentication schemes should be evaluated in the context of fallback authentication to simulate a realistic fallback scenario (see Chapters 5 and 6).

1.4 Thesis Overview

This thesis is organized in eight chapters that, in turn, can be structured in three main parts: The first part focuses on the motivation of fallback authentication and creates its basic foundation (Chapters 2, 3 and 4). The second part analyzes different types of autobiographical memories through example implementations that take into account different design options (Chapters 5 and 6). The third part assembles the individual results from each project to create the big picture and to conclude this thesis (Chapters 7 and 8). The content of each chapter can be summarized as follows:

Chapter 2 gives an overview of existing work that is related to fallback authentication. In particular, we motivate the need for fallback authentication by describing the memorability issues of passwords in general and by discussing the different measures that are available to address these problems. We further provide an overview of current fallback solutions in practice and depict their shortcomings, focusing on the use of security questions, which have also found some interest in the research community. Since this thesis sets its focus on smartphones, we present an overview of different types of mobile lock screen authentication to identify potential issues that may lead to lockout situations. Due to the lack of research on mobile fallback authentication, we review how smartphones are used to support fallback authentication instead. The chapter concludes with an overview of the human memory and the different memory structures, with a focus on autobiographical memories to motivate their use for the design of fallback schemes.

Chapter 3 provides detailed insights into the frequencies of lockout experiences as well as the problems and countermeasures that users encounter in these situations. We present the results of an online survey and complementary interviews on this matter. The chapter provides a basic understanding for mobile fallback authentication and identifies important requirements for the design of fallback schemes in a mobile context.

Chapter 4 derives a framework for fallback authentication that is based on the insights from the previous chapters. The framework includes a clear definition of fallback authentication and identifies the properties that distinguish it from primary authentication. The properties are then translated into requirements and inspire the design of each prototype presented in this work. In addition to this, the chapter overviews the design space of fallback authentication and presents our research methodology.

Chapter 5 presents the design and evaluation of two fallback schemes that are based on static enrollments. While one of the schemes focuses on the skill to draw with two fingers simultaneously, the other uses episodic memories with a strong spatio-temporal context for the creation of location-based questions. We summarize the results to point out the potential and challenges of the two memory types and make suggestions for improvements. The chapter further highlights the importance of different usability factors that were assumed to be less relevant in the context of fallback authentication.

Chapter 6 complements the preceding chapter by focusing on dynamic enrollments. We exploit different types of smartphone activities to implement three example schemes for fallback authentication. In particular, our implementations set a focus on app-related activities due to their promising trade-off between memorability and security. In addition, the chapter identifies further factors that need to be taken into account for the design of dynamic enrollments, such as privacy and data availability. Therefore, these factors extend the list of factors that have been identified in Chapter 4.

Chapter 7 summarizes the key insights from the design and evaluation of each individual implementation. Its main objective is to motivate the importance of fallback authentication in research despite its infrequent occurrence as well as to overview the implications from the study results for the design and evaluation of fallback schemes in general. We further make recommendations for the use of autobiographical memories in the context of fallback authentication.

Chapter 8 concludes this thesis with a summary of the main contributions and presents our final definition of fallback authentication that is inspired by the study insights. We further discuss the limitations of this work to give an outlook on the vision that we have for research in the field of fallback authentication.

2

Learning from the Past and Present

The use of passwords for authentication is a common procedure deployed by many web services and is unlikely to change in the near future. Yet, they come with drawbacks like the preference of users to select simple passwords due to memorability reasons. Despite the efforts made to influence password selection, users often still pick easy-to-guess passwords, perhaps due to the inconvenience of password loss (Section 2.1). Fallback schemes for password recovery, such as the use of security questions, have shown major usability and security issues. In particular, the memorability of security questions has to be considered critically due to their infrequent use (Sections 2.2 and 2.3).

Password selection is aggravated on mobile devices (e.g., smartphones) due to their small form factor. Therefore, alternative schemes for primary authentication on these devices have been proposed. However, users seem reluctant to adopt them due to the needed training. Especially in the early phases users are prone to errors and password loss (Section 2.4). However, documentation on password recovery when lockout happens on mobile devices is almost non-existent. This makes recovery difficult for inexperienced users (Section 2.5).

This highlights the need for alternative fallback schemes, which, however, require further understanding of human memory to overcome the memorability issues of current solutions (Section 2.6). The human memory is a complex structure, consisting of multiple memory types. In particular, autobiographical memory seems to be a promising source for the design of fallback authentication schemes (Section 2.7).

2.1 Passwords, the Good and Evil

In this section we introduce the problems of user-chosen passwords and discuss different practices to address these problems.

Introduced in the mid-sixties [34], passwords are still prevalent in the digital era due to their convenience of use [84, 86]. While the number of passwords was manageable in the early days, their number has increased steadily [180]. It is estimated that users have between 8 and 25 distinct accounts that require login (e.g., [37, 64]). In contrast to this, the number of distinct passwords that they use to protect these accounts is few [37, 56, 64, 83, 166]. Still, remembering them all remains a challenging task. In a study by Chiasson et al. users could recall only 70% of their passwords over a retention period of two weeks [27]. It is not surprising that users start to develop undesirable habits like selecting weak passwords, reusing them across accounts or writing them down [3, 15, 37, 83, 166, 180]. Typically weak passwords are kept simple, short and are often based on biographical data like names or date of birth [114, 146, 166, 180]. These kinds of passwords are vulnerable to brute force attacks, dictionary attacks and other guessing strategies [37, 125, 184]. Therefore, websites have put some effort in encouraging users to create stronger passwords.

2.1.1 Password Advice

One approach is the use of password advice to support users during password creation (Figure 2.1, left). However, less than 25% of Internet websites offer this kind of support [17] and the recommendations found there are of different quality. Some of them are vague, outdated [58, 86] or too onerous [117]. In particular the latter situation impacts the memorability of passwords or leads users to ignore given advice. Users do not do this because they are lazy, but mostly due to the imbalance between the required effort and the perceived benefit of following advice [85]. Therefore, it is important to keep password advice short to minimize the effort for users. For example, this can be achieved by neglecting threats that exist in theory, but that are unlikely to occur in real life [58].

2.1.2 Password Policies

A more stringent variant of password advice is the use of password policies that enforce the compliance to certain requirements [188] (Figure 2.1, center). These requirements do not follow a common standard, but vary across websites, ranging from loose policies (e.g., allowing even one-character passwords) to stricter ones (e.g., passwords of at least eight characters from multiple character sets) [57]. However, stricter password policies do not automatically lead to more secure passwords. Therefore, their selection needs to be done carefully [111, 182]. Policies that are too strict impose a high burden on users, decreasing the memorability of the selected passwords [93, 159]. They also increase user frustration and

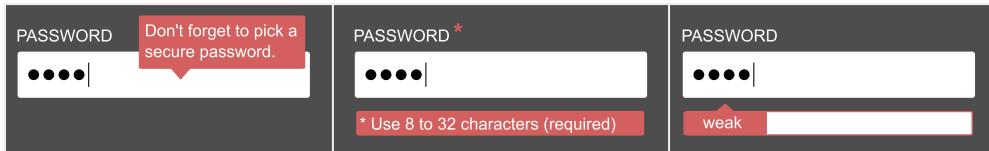


Figure 2.1: Three types of support during password creation: password advice (left), password policy (center), password meter (right).

cause the selection of passwords that fulfill the minimal requirements [120]. Studies by Das et al. and Shay et al. showed that users seldom create new passwords, but prefer to use old ones to adapt them accordingly, for example, by attaching digits at the beginning or end of the password [37, 159].

2.1.3 Password Meters

Another recommendatory approach is the use of password meters that display the strength of a password (Figure 2.1, right). These representations are based on so-called scoring algorithms that, in turn, can be based on different approaches (e.g., [24, 162]). Popular implementations found on the web take password length, the use of numbers or blacklisted words as a reference for strength calculation [177]. However, many of these calculations are inaccurate, accepting weak passwords as strong and the other way around [183]. In addition to the different scoring algorithms, password meters can also differ in their representation. This includes text-based variants, bar-like visualizations or checkmark indicators [177]. Fifteen different representations were tested by Ur et al. and their results showed that any kind of visual representation encourages users to create longer passwords [177]. However, most users achieve this by attaching additional characters, digits or symbols at the end of their chosen passwords. This indicates that users seldom create new passwords and the influence of password meters is limited [180]. Egelman et al. argue that their influence is highly context dependent, such as the importance of the account to be protected [53].

2.1.4 Short Summary

As can be seen from all this work, a lot of approaches have been proposed to motivate users to select stronger passwords. Yet, the silver bullet has not been found as password advice, password policies or password meters are often ignored by users. One of the reasons is that password recovery, in particular in companies, is often cumbersome [93]. But even if one of these mechanisms succeeded, passwords would not become more or less memorable. Therefore, the availability of usable and secure recovery mechanisms that users can fall back on remain important in two ways: First, they are a safety net to rely on when all other means fail. Second, they may inspire users to follow security advice to create stronger passwords when easy password recovery or reset is available in case of password loss.

2.2 Current Practices in Fallback Authentication

In order to get a grasp of available fallback schemes in practice and their shortcomings, this section summarizes corresponding insights from previous research and further presents the results from an informal analysis.

Due to the many accounts that users have to manage, it is inevitable that, at some point, user-names or passwords are forgotten [64]. Therefore, fallback solutions are needed to enable users to regain access to their accounts and data [64, 122]. It is assumed that this happens once per month or less [166]. In order to use fallback schemes in lockout situations, users are required to set them up beforehand, for example, during registration. For this, users often can choose from a set of fallback schemes. The selection, however, is very limited and has hardly changed over the past few years. Research in this area suggests that users prefer the use of security questions and email-based password resets [94, 122].

2.2.1 Fallback Authentication between 2007 and 2009

An assessment of password practices in 2007 showed that common mechanisms for fallback authentication included the use of email links or security questions [60]. The latter was also often found for online banking websites in 2008 [139] and for popular webmail providers in 2009 (i.e., Google, Yahoo, AOL and Microsoft) [154]. While security questions are no longer in use by Google and Microsoft, they have been kept by Yahoo and AOL as complementary fallback options to email-based password resets. Interestingly, Yahoo still offers the question “*Where did you meet your spouse?*”, although it has shown to be vulnerable to educated guesses, in the case of U.S. Governor Sarah Palin, whose account was hacked in 2008. Weak security questions were also the reason for the success of recent attacks on celebrity accounts for Apple’s iCloud Service that led to the theft of sensitive photos.¹

However, not only celebrities are prone to these types of attacks. A non-scientific experiment by artist Risa Puno showed that people are willing to exchange their personal data (e.g., fingerprint, date of birth or mother’s maiden name) for a single cookie, revealing information that is often required to answer security questions.² These observations are in line with previous research that showed the willingness of users to trade personal information for small monetary amounts (e.g., [69, 79, 92]). Hann et al. tested whether monetary reward and convenience affect user preferences for websites with different privacy policies. They found that when the offered reward exceeds a threshold of 20\$, users are willing to lay aside their privacy concerns [79].

¹ <http://www.apple.com/pr/library/2014/09/02Apple-Media-Advisory.html> (*last accessed 26/12/2014*)

² <http://www.propublica.org/article/how-much-of-your-data-would-you-trade-for-a-free-cookie> (*last accessed 26/12/2014*)

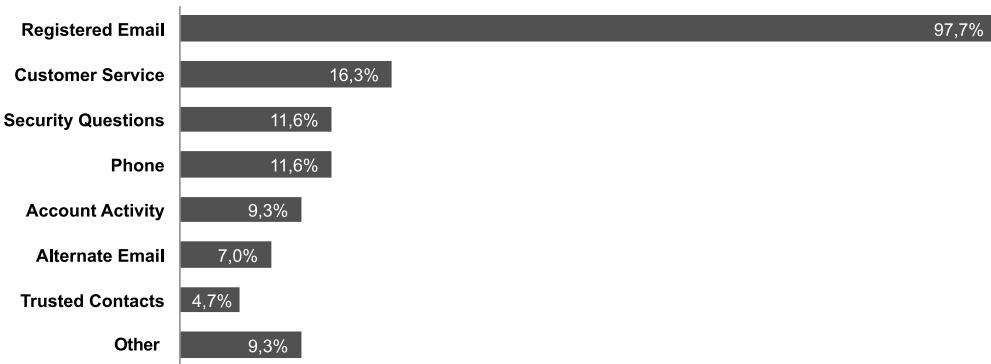


Figure 2.2: The distribution of fallback authentication schemes of Alexa’s top websites in 2014.

2.2.2 Fallback Authentication in 2014

Though still available, the use of security questions has diminished over the past few years. In December 2014, we analyzed the available fallback options of 43 websites after removing duplicates and no-signup websites from the 100 top websites (as listed by Alexa³). Figure 2.2 provides an overview of the distribution of the various fallback schemes that we discovered after consulting the support pages of the corresponding websites and visiting their forgot-my-password links. The majority of websites relies on email-based password resets (97.7%). Only few of them have other fallback options, such as security questions (11.6%), phone-based password resets (11.6%) or trusted contacts (4.7%). Instead, most of them refer their users to customer service when email-based password reset does not work (16.3%). Other fallback schemes were the use of alternate (i.e., not previously registered) email addresses for password recovery (7%). This approach was mostly used in combination with the provision of account-related information to verify the identity of the user (4.7%). A detailed overview of each website and its fallback scheme can be found in Appendix A.1.

2.2.3 Password Managers

Password managers support users in storing and organizing their passwords (e.g., LastPass⁴) [166]. In case users forget their password to one of their accounts, they can use the password manager to look it up. In order to do so, they usually have to authenticate with a master password to get access to the password manager. This is, although not in the traditional sense, another form of fallback authentication. However, this also means that password managers are a single point of failure. Once the master password is revealed, potential adversaries get access to many accounts. In general, the use of password managers is helpful, but needs further improvements in terms of usability and security (e.g., [28, 106]).

³ <http://www.alexa.com/topsites> (*last accessed 29/12/2014*)

⁴ <http://www.lastpass.com> (*last accessed 27/12/2014*)

2.2.4 Short Summary

All things considered, email-based password recovery/reset has been, and still is, one of the most common procedure used for fallback authentication. Though this approach works well [63], it does not take into account situations in which email access is not possible (e.g., expired accounts or public kiosks that the user does not want to trust with the email password). Only few web services offer alternative options. One of the reasons may be the implementation costs, but also the increased security risks with each additional authentication option that give potential adversaries the chance to circumvent the primary authentication.

2.3 The Design of Security Questions

Most research on fallback authentication is centered around security questions. The insights from this research are discussed in this section to depict their potential and challenges. Research on security questions can be traced back to the late eighties when Smith proposed the use of word associations for authentication that, in the broad sense, can be considered as a special kind of security question [161]. His approach was based on passwords that consisted of multiple cue-association pairs (e.g., pink-grapefruit, water-melon). For authentication, the cues (e.g., pink) were shown to ask users about the associations they provided when setting up the password (e.g., grapefruit). Zviran and Haga extended this idea and suggested the use of cognitive passwords: security questions that are based on personal facts or opinions of the user (e.g., “*What is your mother’s maiden name?*” or “*What is your favorite color?*”) [190]. Interestingly, most of the proposed questions have been adopted by many web services for automated password reset.

2.3.1 Classification

While Zviran and Haga differentiate between fact-based and opinion-based questions [190], the design of security questions can be influenced by additional factors. In general, one can distinguish between three types of questions and three types of answers [101].

Question Types

Security questions can be fixed, open and controlled [55, 101, 132, 133]. Fixed questions are predefined and do not allow any customization. Usually, they are presented in a list from which users can choose one or more questions that are found to be most applicable [101]. Typical examples are shown in Table 2.1.

Open questions, in turn, leave room for creativity and require users to define their own questions as free text. Such questions are, at least in theory, highly individual [101]. However, studies have shown that users often lack creativity and define questions with low entropy answers [103].

AOL	Yahoo
<i>What was the name of your first pet?</i>	<i>In which city did you study abroad?</i>
<i>What was your childhood nickname?</i>	<i>What was your first holiday destination ever?</i>
<i>What was your favorite childhood book?</i>	<i>Who was your date on prom night?</i>
<i>What was the name of your first school?</i>	<i>Where did you spend your honeymoon?</i>
<i>In which city did your parents meet?</i>	<i>Where did you meet your spouse?</i>
	<i>What is your oldest cousin's name?</i>
	<i>What is your youngest child's nickname?</i>
	<i>What is the first name of your oldest [niece nephew]?</i>
	<i>What is the first name of your favorite [aunt uncle]?</i>
	<i>What town was your [mother father] born in?</i>

Table 2.1: Subset of the security questions used by AOL and Yahoo in 2014.

The third question type, controlled questions, is a hybrid of fixed and open questions that occurs in different variations. One of them uses variable fragments that are indicated by blanks [101], such as: “*What is your favorite ___ ?*”. The blanks can be filled by the user with different terms, such as “*food*”, “*color*” or “*restaurant*”. Another variant is the definition of hints for each question to remind users of their answer during authentication. For example, Renaud and Just propose a system in which users select an image per question to serve as an associative cue [144]. However, it has been shown that customization is rarely done and rather targeted at more advanced users [122].

Answer Types

Analogous to question types, answers to security questions can be fixed, open or controlled. Fixed answers present to users a variety of options for selection. This includes, for example, drop-down lists or multiple choice fields. In turn, open answers require users to enter free text (usually in a text field). A special form of open answers is the use of controlled answers that impose further restrictions, such as the compliance to a certain format (e.g., dates that have to be entered in the form of *YYYY-MM-DD*) [101].

2.3.2 The Usability of Security Questions

Although the various question and answer types provide some combination possibilities for the design of security questions (e.g., [132]), each of them comes with potential usability and security issues. For example, fixed questions do not allow any kind of customization and thus can cause applicability issues. Open questions, in turn, require creativity that not all users may possess [55, 100, 103]. In general, there are three major usability weaknesses: inapplicability, ambiguity and low memorability [100, 139].

Inapplicability

A question that most users are not able to answer truthfully is inapplicable [139]. For example, the question “*What was your first pet’s name?*” does not apply to a large fraction of the public as about 40% of the U.S. population does not own a pet (as of 2012).⁵ While it is acceptable if some questions within a set are inapplicable, the problem becomes severe when their number increases as it limits the choices that users have during enrollment. An analysis of security questions from 16 online banking websites in 2008 identified almost 50% of the questions to be inapplicable. Most of those questions referred to spouses, marriage or children [139]. Interestingly, these are exactly the types of questions that are still in use by Yahoo (Table 2.1, right). The lack of viable options is critical as it may cause undesirable behaviors, such as choosing less memorable questions (that are at least applicable) or answering questions untruthfully [154]. Both behaviors can lead to memorability issues.

Ambiguity

There are two kinds of ambiguity when it comes to security questions: semantic ambiguity and lexical ambiguity [100]. The former is of concern when there exists more than one possible answer to a question. For example, when asked for their favorite actor, users may consider more than one person as a possibility. When the actual fallback authentication takes place, users may not remember which answers they decided to provide during enrollment. It is also possible that their favorite actor has changed over time [139].

Lexical ambiguity, in turn, refers to the variations in the representation of an answer [100]. For example, though semantically similar, the term “*street*” can include uppercase letters (e.g., “*Street*”, “*STREET*”), use lowercase letters only (e.g., “*street*”) or use abbreviations (e.g., “*St*”) [59, 154]. An analysis of 117 security questions by Just and Aspinall revealed that after four weeks 18% of the answers were semantically correct, but did not match exactly the answers provided during enrollment [102]. Most fallback systems are tolerant to small lexical variations. However, there exist also more complex variations that are more difficult to implement [59]. In those cases, users are often left with no other choice but to walk through numerous possibilities until they find the correct one or until they have exhausted the number of allowed attempts [154]. In particular, ambiguity is of relevance when open or controlled questions are implemented. The problem does not exist for fixed answers where users select from a set of options [100]. However, using fixed answers comes with other issues like facilitated guessability [101].

Memorability

Memorability refers to the ability of users to recall the answers to security questions. As those questions are based on personal information, it is assumed that no explicit memorization is required by the user so that recall is made easier [190]. In reality, memorability

⁵ <https://www.avma.org/KB/Resources/Statistics/Pages/Market-research-statistics-US-pet-ownership.aspx>
(last accessed 05/11/2014)

First Author	Year	Respondents	Time	Question Type	Recall	Guess
Smith [161]	1987	4	6 months	Association	94%	n/a
			18 months	Association	86%	n/a
Zivran [190]	1990	106	3 months	Opinion	88%	23%
				Fact	94%	37%
Haga [72]	1991	106	3 months	Association	69%	26%
				Opinion	70%	33%
				Fact	84%	45%
Zivran [191]	1993	103	3 months	Association	69%	n/a
				Opinion	84%	n/a
				Fact	74%	n/a
Podd [136]	1996	86	2 weeks	Association	39%	7%
				Opinion	72%	23%
				Fact	88%	56%
Bunnell [23]	1997	90	2 weeks	Association	39%	7%
				Opinion	72%	23%
				Fact	88%	56%
Pond [138]	2000	73	2 weeks	Association	66%	12%
Just [102]	2009	39	28 days	User Chosen	75%	n/a
Just [103]	2009	97	23 days	Mixed	8%	n/a
Schechter [154]	2009	130	6 months	Mixed	80%	22%
Renaud [144]	2010	90	1 week	Mixed	56%	32%

Table 2.2: Overview of recall and guessing rates of security questions from different research. The table includes results about association-based, opinion-based, fact-based and user-chosen security questions. Some studies did not distinguish between the different question types and are depicted by the term *mixed*.

issues are not uncommon, in particular when the time between enrollment and authentication increases. While early studies showed that users were able to recall almost 95% of their answers after several months [72, 190], these numbers have to be interpreted carefully (Table 2.2). Those studies were either with very few participants or came with high guessability rates. More recent studies (between 2000 and 2010) showed lower recall rates that ranged between 56% [154] and 80% [102]. Table 2.2 compares the different recall rates from different user studies.

2.3.3 The Security of Security Questions

While usability factors focus on the ability to recall the answers to security questions, it is equally important to analyze such questions from a security point of view. In general, there are three factors that influence the security of security questions: guessability, researchability and the type of adversary [101, 139].

Guessability

The term guessability refers to the difficulty with which an answer to a security question can be guessed without any knowledge about the user [101, 139]. Many questions have a small answer space [94, 101]. For example, the question “*What is your eye color?*” is likely to be guessed within a small number of attempts [101]. A security analysis by Rabkin [139] classified at least 33% of security questions from 16 different financial websites as guessable, meaning that they found an answer for these questions that was likely to be correct in excess of 1%. Another analysis by Schechter et al. showed that 13% of security questions could be answered through statistical guessing by choosing the five most popular answers to the corresponding question [154]. In particular, questions about names have proven to be vulnerable to statistical attacks, due to the lack of diversity of human names [16].

Researchability

Many security questions can be attacked by additional knowledge about the user [139]. This knowledge can be obtained by personal communication, online research or even automated processes that, for example, mine the needed information from online resources, such as search engines, personal homepages or social networks [68, 101, 139, 140, 163]. This also means that the success of such attacks depends strongly on the availability of information. Research has shown that social networks have become popular platforms for self disclosure [99, 112] that reveal many personal information like contact details (e.g., email, instant messaging or full address [88, 110]), personal interests (e.g., hobbies, political orientation and sexual orientation [48]) and other personal data (e.g., date of birth [2]).

Type of Adversary

The success of such attacks (guessing as well as researching) depends on the type of adversaries. Attacks can be automated (formed by a computer) or by dedicated human beings [139, 161]. It was early noticed that adversaries that are close to the user (e.g., significant other) are a serious threat. They, in comparison to strangers, know much more about the user and thus are also more likely to know the answer to a question [72, 101, 122, 190]. Table 2.2 lists the guessing rates from various user studies that have been conducted over the past years. One of the key insights from the early studies was that fact-based security questions, though easy to be recalled by users, are also more easily guessed by close adversaries than any other type of question [23, 136, 190].

2.3.4 Advancements in Fallback Solutions

Since most security questions lack in usability, security or both, some effort has been made to improve them or to find alternative approaches for them. Some of these solutions are discussed next.

Social Authentication

For example, some proposals include the help of third persons, such as friends or acquaintances, that are appointed by the user during registration [19, 155]. In case of lockout, users are required to contact a subset of them (e.g., by mail or phone) to collect codes that are needed for account recovery [155, 158]. Though the approach works well, it still suffers some usability and security issues [155]. Recalling the trustees one has appointed is difficult when a long time between registration and re-authentication has passed. In general, the selection of trustees has to be done carefully for multiple reasons. Relationships may change over time and some trustees may not be as trusted anymore. It has also been shown that trustees that are not so close to the user are prone to revealing recovery codes through phone-based attacks.

Another form of social authentication is the use of photos from social networks. In order to authenticate, users have to answer a number of questions, such as identifying persons on these photos [189]. Due to the use of multiple questions (and thus the time needed for authentication), the approach seems to be most suitable as a second factor of authentication⁶ or as fallback option. However, social authentication has to be viewed critically as it comes with certain drawbacks. Advances in the area of face recognition make the threat of automated attacks more prevalent. Human adversaries are often in a close relationship to the user (e.g., jealous spouse) and have advanced knowledge or access to information (e.g., common friends) [109]. But also strangers can research related information by issuing friend requests and browsing friend lists in social networks [137].

Preference-Based Security Questions

The basic idea of preference-based security questions is inspired by online dating platforms that allow users to state their likes and dislikes on certain topics in order to find potential matches. In the context of fallback authentication, users are presented with a set of preference-related questions, such as “*Do you like country music?*”. The answers are provided by selecting from three answer options (i.e., really like, really dislike and neutral) [97]. In a simplified version of the described approach, users are presented with a set of images from which they have to choose five items that they like and dislike, respectively. The remaining images are classified as neutral [96].

It is assumed that preferences remain stable over time, meaning that radical preference shifts (e.g., from really like to really dislike) happen less frequent than small preference changes (e.g., from really like to neutral) [97]. Though preference-based authentication yields promising results in terms of false positives and false negatives [96, 97], they come with further shortcomings, such as the high number of questions required [97], insider threats [96] or the mining of preferences from social networks [70]. Furthermore, the stability of preferences has to be considered critically as there does not exist a clear consensus on this topic [90].

⁶ <https://www.facebook.com/notes/facebook/a-continued-commitment-to-security/486790652130> (last accessed 18/12/2014)

Image Priming and Labeling

Another image-based approach primes users during enrollment by showing them a set of images that they have to label. This procedure is repeated during authentication, but with additional images that have not been shown before. It is assumed that priming affects the labeling performance of users, in such way, that primed images are more likely to be labeled correctly than non-primed images. Though the results seem promising, the priming effect is not as strong as expected [46].

Context-Based Security Questions

Some attempts have been made to design security questions that take the user's current context into account. This can be trivial information from the user's environment, such as "*Where is a wall clock in your house?*" [132, 133], but also includes more dynamic information like office arrival/departure or recent events from calendars (e.g., [130, 131]). The use of context information has the potential of overcoming issues like applicability as well as memorability. However, these types of questions have gained little attention by the usable security and privacy community and thus will be explored more thoroughly in this thesis.

2.3.5 Short Summary

All these examples show that the design of security questions is a challenging task. While the perfect security question is easy to remember by the legitimate user and hard to guess by potential adversaries, this requirement seems impossible to be met. Current security questions come with usability issues (i.e., inapplicability, ambiguity, and memorability) as well as security threats (i.e., guessability and researchability) by different types of adversaries. Though several alternatives have been proposed to fact-based security questions, the best trade-off between usability and security has yet to be found. It is clear that any fallback option will have some negative and some positive aspects. But in particular for a mobile context, the discussed solutions do not seem to be a good fit. For example, social authentication is difficult in a mobile context when the device is locked and contact details on the phone are not accessible. Furthermore, fallback authentication on mobile devices happens infrequently (see Chapter 3) so that preference-based approaches are problematic due to possible preference changes. So far, dynamic fallback approaches have shown promising results, but have not yet been evaluated thoroughly. One part of this thesis picks up this aspect and contributes to deeper insights in this area.

2.4 Smartphone Authentication

This section gives an overview of primary authentication schemes on smartphones and their problems to motivate the need of fallback authentication.

While the design of passwords and their corresponding fallback options has mostly focused on the desktop environment, their importance in the mobile environment has increased due to the transformation of mobile devices from simple communication tools to so-called Swiss army knives that contain a lot of sensitive data [13, 151]. However, the small form factor and keyboard characteristics of smartphones make it difficult to enter textual passwords (as used in desktop environments) [153]. This has aggravated the situation of password selection: Users prefer even shorter passwords [181]. Thus, several alternative authentication methods are available on smartphones, such as personal identification numbers (PINs), graphical passwords or biometric approaches.

2.4.1 Passwords and PINs

PINs on smartphones often consist of a sequence of only four digits, since most users have difficulties in memorizing more complex ones [53]. Nonetheless, users have many memorizing strategies that come at the cost of security [9]. An analysis of real PINs from over 200000 iPhone users showed that even short PINs are kept as simple as possible (e.g., 0000 or 1234).⁷ They are often based on birth dates [18] or consist of numbers that are located close to each other [108]. In the last resort, when even these questionable methods fail, users start to write down their PINs [141].

2.4.2 Graphical Passwords

In recent years, graphical passwords have become a popular authentication method on Android devices. Their idea was first introduced by Blonder⁸ and was based on the pictorial superiority effect, which assumes the better memorization of visual input [129]. In order to authenticate, users are required to draw their password on a given grid (e.g., [43,98,180,185]) or background image [50]. While this approach is based on pure recall and often referred to as drawmetrics, there exist other variations that use cued-recall (i.e., locimetrics) or recognition (i.e., searchmetrics or cognometrics) [12, 38, 143, 186]. Popular examples for cognometrics are Use Your Illusion [82], DéjaVu [47] or PassFaces [20], while PassPoints [186] or Cued Click Points [29] are representatives of locimetrics. Though it is easier for users to memorize graphical passwords than PINs [123], graphical authentication systems have similar problems [12, 124]. While the theoretical password space may be large, users tend to select the obvious [49, 142] due to memorability reasons [38, 142, 143, 186].

⁷ <http://danielamitay.com/blog/2011/6/13/most-common-iphone-passcodes>

⁸ Patent: Blonder, Greg E. Graphical password. U.S. Patent No. 5,559,961. 24 Sep. 1996.

2.4.3 Biometric Passwords

In general, one can distinguish between physiological biometrics (e.g., fingerprint, face or iris) and behavioral biometrics (e.g., gait or location [35, 61, 172]) [187].

Popular implementations that use physiological features are Apple's Touch ID⁹ or Android's Face Unlock¹⁰. While both systems take advantage of only one feature (i.e., fingerprint and face, respectively), a combination of multiple features is also possible for biometric authentication (e.g., [21, 26, 91, 118, 135, 147]).

In contrast to this, behavioral biometrics authenticate users implicitly by the way they do certain things (e.g., [40]). Popular behavioral cues are the use of gait (e.g., [35, 61, 172]), location (e.g., [95, 160, 172]) or keystroke dynamics (e.g., [22, 113]).

Though the use of biometric systems has gained popularity, there are still circumstances in which authentication does not work. Dirty fingers can make fingerprint authentication difficult, while bad lighting makes Face Unlock impossible. In these cases, alternative means of authentication are needed [41]. The same applies for behavioral cues that are often proposed for continuous authentication. In case users behave differently than expected by the system, they have to authenticate through other means, such as PINs or passwords (e.g., [172]).

2.4.4 Short Summary

There exist a variety of knowledge-based authentication schemes to protect sensitive data on smartphones (e.g., PINs or graphical passwords). Still, a remarkable number of smartphone owners do not protect their device [52] and those who do, exhibit insecure behaviors by selecting weak passwords [181]. Thus, some effort has been made to design alternative systems that encourage the selection of stronger passwords on mobile devices. However, the need for fallback options remains as users may still forget their passwords, for example, when a new authentication scheme is set up [20, 38, 47, 50, 186].

In contrast to this, biometric authentication schemes do not have any memorability issues. Nonetheless, it is important to understand that in case biometric features are compromised, it is difficult (in most cases impossible) to change them. Furthermore, false negatives are inevitable when using biometric authentication and thus the need for knowledge-based (or token-based) alternatives remains. The design of those alternatives is challenging. The better a biometric system works, the rarer these alternatives are needed, causing potential recall issues when the actual fallback authentication takes place.

⁹ <http://support.apple.com/en-us/HT5883> (*last accessed 27/12/2014*)

¹⁰ <https://support.google.com/nexus/answer/2781894?hl=en> (*last accessed 25/10/2015*)

2.5 Fallback Authentication on Smartphones

The rare occurrence of fallback authentication on smartphones is reflected in the little interest that it has received, until now, by smartphone manufacturers and the research community. It is unclear what the requirements of mobile fallback authentication are and whether current desktop solutions can also be applied in the mobile context.

Research in the area of mobile fallback authentication has mainly focused on the use of mobile devices as assisting tools for authentication at public terminals (e.g., [39, 42, 157]) or for password recovery. A common example for the latter is the use of mobile devices for phone-based password resets.¹¹ In case of lockout, a temporary password is sent to the user's mobile phone number to log into the blocked account and to create a new password. Mannan et al. extend this idea with a public-private key infrastructure that is used to encrypt and decrypt the new password, respectively [116].

In general, the documentation of actual fallback authentication on smartphones is lacking. This thesis will cover this neglected area of research to provide deeper insights into mobile fallback authentication, including the circumstances in which lockout happens and the requirements that need to be taken into account (e.g., memorability factors). This will help to understand how fallback solutions should be designed and whether current desktop solutions could also be applied to the mobile context.

2.6 The Human Memory

Since memorability has been depicted as one of the key factors of current fallback solutions, it is essential to learn about the general structure of human memory. Understanding the functions, strengths and weaknesses of the different memory types will be helpful to guide us in the design of alternative fallback schemes.

2.6.1 The Multi-Store Model of Memory

The human memory was long treated as a unitary system [7]. However, this view started to degrade in the late sixties when Atkinson and Shiffrin proposed their multi-store model of memory which assumes the existence of three, closely related, memory systems: the sensory register, the short-term store and the long-term store [5]. Figure 2.3 depicts the three memory types and their relation to each other. The sensory register is responsible for capturing stimuli from the environment. While most of this stimuli decay within a fraction of a second, a small subset of it is transferred to the short-term store for further processing (e.g., by complementing the input with previous knowledge). The input is lost when not rehearsed or not transferred to a more permanent structure (i.e., the long-term store) [7].

¹¹O'Connell, Ellen R. Automated Password Reset. U.S. Patent No 5,991,882, Nov 1999

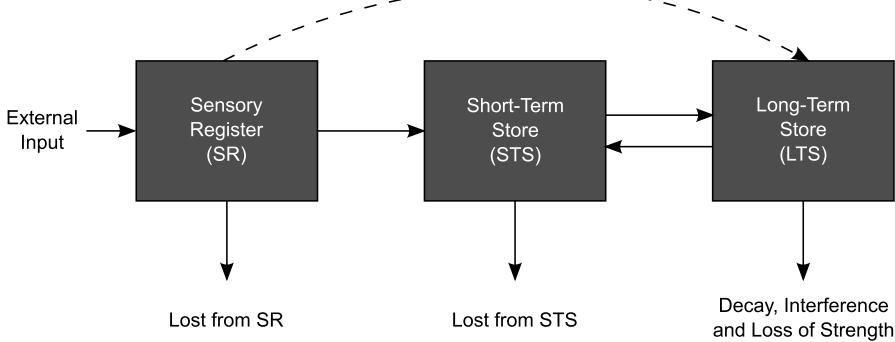


Figure 2.3: Structure of human memory by Atkinson and Shiffrin. Reproduced from [5].

The three memory components correspond to the main tasks of the human memory: *encoding* (i.e., the registration of stimuli by the sensory register), *storage* (i.e., the maintenance of information in the long-term store) and *retrieval* (i.e., the recall of memories from the long-term store) [8]. The latter task can be further divided into two modes that Ebbinghaus refers to as voluntary and involuntary retrieval [51]. While voluntary retrieval happens by the use of the will, the involuntary retrieval is triggered by associative cues [11, 51].

2.6.2 Long-Term Memory

From the three presented memory stores, the long-term store is the most important one for fallback authentication as it is the basis of an individual's complete knowledge. Long-term memories can be of explicit or implicit nature and encompass knowledge of different kinds [165] (Figure 2.4). Implicit memories, for example, refer to performance-related knowledge like motor skills or habits. Explicit memories are further distinguished in semantic and episodic memories. The main difference between the two lies in the information they hold and the way they are accessed [175, 176]. While semantic memories consist of general knowledge about the world and its relation to other knowledge, episodic memories relate to personal experiences with spatio-temporal context. We often refer to them as something we *know* and something we *remember*, respectively.

Imagine, for example, a girl that had her first kiss at the banks of the Seine when she was sixteen. This experience is stored in episodic memory and can be *remembered* to relive the past. In turn, *knowing* that the Seine is a river in France is semantic information that is detached from personal experiences, as temporal as well as spatial information are absent.

Tulving describes this phenomenon as anoetrical retrieval (for semantic memories) and auto-noetrical retrieval (for episodic memories) [176]. This distinction is similar to Ebbinghaus' classification of voluntary and involuntary retrieval [51].

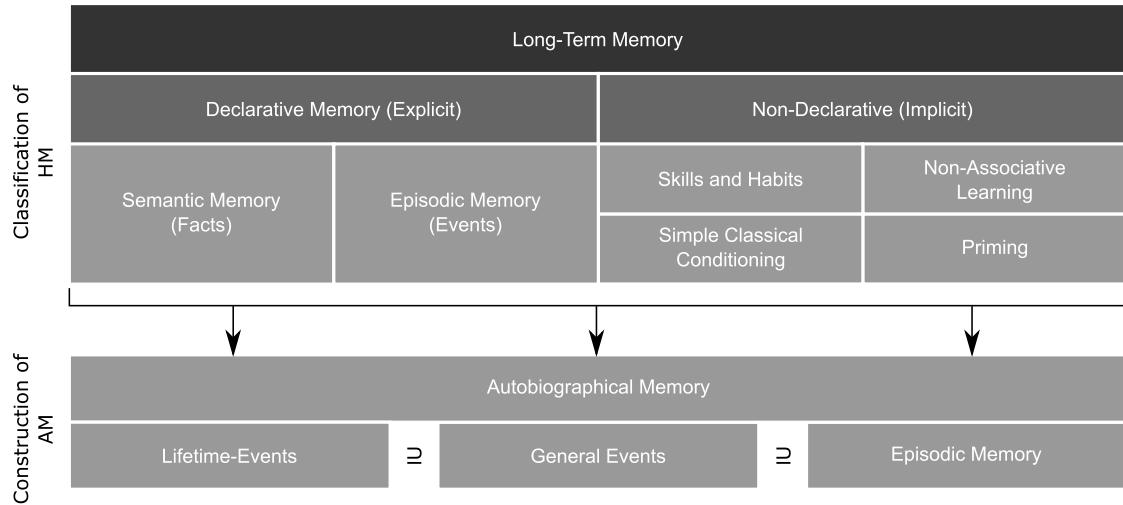


Figure 2.4: Classification of human memory (HM) and construction of autobiographical memory (AM). Reproduced and adapted from [165].

2.6.3 Autobiographical Memory

The term “*episodic memory*”, as introduced in the previous section, has often been (and still is) treated synonymously with the term “*autobiographical memory*”. Even though both types of memories are closely related, more recent research has attempted to find a clearer distinction [32]. It is assumed that the retention times of episodic memories are very short and in the dimension of minutes, hours or days, while autobiographical memories are stored for days, weeks or even for a lifetime [33]. Nonetheless, episodic memories (not all) can become part of autobiographical memory when a transfer from one to the other is initiated. Such transfer depends on various factors like importance, involvement, repetition and presentation of the experienced episode [119].

Though autobiographical memory consists of many episodic memories, it also includes other types of information that come from the semantic memory (e.g., one’s date of birth) as well as the procedural memory (e.g., one’s skill to drive) [32,119]. All this information is retained with different levels of specificity and forms a hierarchical structure that includes life-time events, general events and event-specific episodes [32,33]. Consider the following example: “*When I was in high school, I liked walking in the woods. However, during one of these walks, I stumbled and broke my leg.*” These memories refer to a life-time period (e.g., the time in high school), which in turn consists of general events (e.g., the habit to walk in the woods) that, again, consists of event-specific episodes (e.g., the broken leg).

Based on these insights from psychological research, the remainder of this thesis will define the term “*autobiographical memory*” as follows:

Autobiographical memory reflects an individual's personal life history. It is composed of self-related information from different memory types. This includes personal facts (semantic memories), skills (procedural memory) and episodic experiences (episodic memory). All this information is retained with different levels of specificity that can be classified as life-time periods, general events and event-related information.

2.6.4 Short Summary

Human memory is a complex structure that consists of multiple components to encode, store and retrieve information. Since fallback authentication happens infrequently, explicit memories from the long-term store are a promising information source for their design. Current solutions also take advantage of the long-term store, but mainly focus on personal facts and opinions. However, autobiographical memory is not limited to this kind of information, but instead, also consists of life episodes that, once transferred to autobiographical memory, can last for a lifetime. Psychological research has shown that episodic memories are retrieved autonoetically, which is faster and easier than anoetical retrieval. Therefore, this thesis will explore the potential of episodic memories as part of autobiographical memory for the design of alternative fallback authentication systems and evaluate them in terms of usability as well as security.

2.7 Lessons Learned

This chapter has given an overview of relevant work in the field of authentication as well as fallback authentication (both in desktop and mobile environments). The take-home messages of this chapter can be summarized as follows:

- **Users forget passwords.** Users are bad at memorizing complex passwords and thus prefer to select weak passwords. If this problem already is prevalent for primary authentication, it can be assumed to be even more severe in the context of fallback authentication as it happens less frequently. This factor must be considered when designing fallback schemes.
- **Fallback authentication is not well documented.** Though it is indisputable that fallback authentication happens, there exist little information about the situations in which fallback authentication happens, the available fallback schemes and their usage in the wild. This is in particular the case for mobile devices.
- **One solution is not enough.** While email-based password resets are reasonable, the need for alternatives remains as there are situations in which email accounts are not accessible.

- **Security questions are not the answer (yet).** Security questions have major usability issues (i.e., inapplicability, ambiguity, memorability) and security issues (i.e., guessability and researchability). Nonetheless, they are still deployed by some web services despite the incidents where security questions have been exploited to circumvent primary authentication and to steal sensitive data.
- **Autobiographical memories are more than personal facts.** Autobiographical memory reflects a person's personal history. Though it contains personal facts that are often used for security questions, it is also a repository for many other types of information, such as episodes about one's personal life. These memories seem worth exploring as they are stored more permanently and are accessed differently than personal facts.

3

A Field Study on Mobile Fallback Authentication

Smartphone owners spend a substantial amount of time with authentication tasks, such as unlocking their device [80, 178]. However, little is known about the circumstances and frequencies in which these tasks fail, in such way, that complete lockout happens.

This chapter gives an overview of potential lockout scenarios on smartphones (Section 3.1) and provides first insights into the frequencies, reasons, countermeasures and problems of lockout occurrences (Section 3.2). The results are based on an online survey ($n=244$; Section 3.3) and semi-structured interviews ($n=12$; Section 3.4).

We found that lockouts happen infrequently and seldom lead to severe situations. However, special circumstances make mobile fallback authentication difficult and leave room for improvements through design. These insights are helpful to identify aspects that should be taken into account when designing alternative fallback schemes on mobile devices and inspire the fallback approaches presented in this work (Sections 3.5 and 3.6).

Personal contribution statement: The content of this chapter is based on a bachelor thesis by Victoria Müller [127] that was supervised by the author. Part of this work was published at the Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI 2015), with co-authors Alexander De Luca, Emanuel von Zezschwitz, Manuel Demmler and Heinrich Hussmann [77]. A detailed overview of task responsibilities can be found in the disclaimer.

3.1 Fallback Authentication on Smartphones

In general, lockouts on smartphones may occur whenever authentication is needed, for example, when users authenticate with the cellular network, when users unlock their device or when users access their online accounts. The remainder of this section describes the three authentication types in more detail and provides examples on how they may lead to lockouts.

3.1.1 Cellular Network Access

In order to use general phone functionality (e.g., making calls), smartphone owners have to authenticate with the cellular network. The way authentication is done depends on the network used. Due to historical reasons, multiple network types coexist. Popular examples are *Global System for Mobile Communications* (GSM), its successors *Universal Mobile Telecommunications System* (UMTS) and *Long-Term Evolution* (LTE) as well as *Code Division Multiple Access* (CDMA) networks [44].

GSM, UMTS and LTE use subscriber identity modules (SIM cards) to authenticate their users. In turn, authentication with CDMA networks does not require any additional components. However, there are exceptions: Some devices include slots for CDMA subscriber identity modules (CSIM) [1] or SIM cards to enhance the device's functionality, for example, to support roaming or LTE usage.

Since these modules store sensitive information about the user, they are usually protected with a PIN code [152]. In case a PIN code is used, lockouts may occur, when the number of authentication attempts is exhausted, for example, when users repeatedly enter the incorrect PIN. A popular solution for these kinds of situations is the provision of a personal unblocking code (PUC): an eight-digit number that needs to be entered by the user [148]. This number is usually sent to the user by the service provider or can be looked up online.

3.1.2 Device Unlock

Smartphone users can choose from a variety of authentication schemes to protect their data from unauthorized access (see Section 2.4). While most schemes are knowledge-based (e.g., PIN code or Android pattern), some approaches are based on biometric features (e.g., Touch ID or Face Unlock).

Alternatives for these schemes are needed when users fail to provide the correct password within a limited number of attempts. In practice, these alternatives vary and depend on the primary scheme used. For example, Android users with Face Unlock are required to set up a graphical password as fallback option that, in turn, uses email-based resets for fallback authentication.

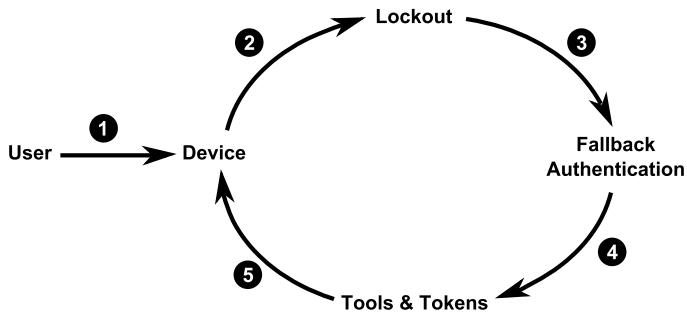


Figure 3.1: Authentication deadlock. (1) User has to authenticate on the smartphone. (2) User fails to provide the correct password and gets locked out completely. (3) User has to rely on fallback authentication. (4) User has to use tools or tokens for fallback authentication. (5) User fails to do so as these tools or tokens are located on the locked device.

3.1.3 Account Login

The use of online accounts is not limited to the desktop environment, but is also commonly done on smartphones by using apps or the mobile browser. Lockouts occur when users forget their credentials. In most cases, there are no fallback schemes that are specifically designed for the mobile context, but instead, common fallback practices known from desktop computers are re-used for this purpose (see Section 2.2).

3.2 Research Strategy

The previous section depicted situations in which lockouts may occur and gave examples of currently available solutions for fallback authentication. The main problem that we see with these solutions is their potential to cause authentication deadlocks, as they often rely on tools or tokens that are not always accessible without the locked device (Figure 3.1). In this section we briefly explain potential problems of current fallback solutions. We further introduce our research strategy to show how we analyze these problems in more detail.

Imagine a user who is prompted to enter the PIN code for the SIM card after an unexpected restart of the device during a vacation. Due to the infrequent use of the PIN code, the user fails authentication and gets locked out completely (Figure 3.1, steps 1-2) and thus needs to provide the PUC for fallback authentication (Figure 3.1, steps 3-4). Since the device is locked, the user does not have the needed Internet access to retrieve it (Figure 3.1, step 5).

Since little data is available on how often these problems occur on smartphones, our research strategy is twofold: (1) we survey a large sample of smartphone owners to get a general overview of the frequency in which lockouts happen; and (2) we complement these numbers with in-depth interviews about lockout experiences to learn about the reasons for lockouts as well as the problems and solutions that users have during these situations.

3.3 General Overview of Fallback Experiences

In order to obtain a general overview of the frequency of fallback experiences, we conducted an online survey ($n=244$) in June 2014 that focused on three lockout scenarios: (1) SIM card lockouts; (2) lock screen lockouts; and (3) account lockouts (see Section 3.1). The design of the survey, its evaluation as well as results are reported next.

3.3.1 Survey Design

Amazon Mechanical Turk (MTurk) was used to conduct the online survey. It is a platform that allows the recruitment of people over the Internet to solve small tasks that require human intelligence. The MTurk platform was shown to be helpful for usable security and privacy research when preventive measures are taken into account (e.g., the use of control questions) [107].

The survey was open to all workers from the U.S. with a HIT approval rate of 90% or higher. In addition to this, workers were required to own a smartphone. Therefore, they were requested to visit a website on their smartphones for device detection. This information was stored in association with the provided worker ID for verification purposes. Smartphone owners who use CDMA devices were not excluded from the survey as some of these devices also have SIM capabilities (see Section 3.1). The survey lasted for about 20 minutes and workers received 1\$ for their participation. The earnings were based on the recommendations provided by Amazon that suggest to pay 4\$-5\$ per hour.

The survey was structured in four parts and started with demographic questions (e.g., gender and age). The remaining parts each focused on one lockout scenario. Each part started with a brief introduction to the lockout type it was dedicated to and was followed by questions about the security measures that participants use on their smartphones (e.g., “*Do you use a PIN code to protect your SIM card?*”).

We also asked participants about their general experiences with [SIM card | lock screen | account] lockouts and asked them to provide more details on their last fallback experience, in case it was available. This included questions about the reasons for lockout, the taken countermeasures and the problems they had to face. In order to avoid that participants answer all questions with *no* (to finish the survey quicker), each branch in the survey consisted of similar amounts of questions. For example, we added authentication-related questions (e.g., “*What do you do to not forget the PIN of your SIM card?*”) to the *no*-branches.

3.3.2 Coding

Two researchers independently developed codes for the free-text answers and met to find an agreement on the codes to create a final code plan that was used by a third researcher.

Age Range	n	%	Highest Education	n	%
18-24	55	22.5	Junior high school	1	0.4
25-34	111	45.5	Senior high school	51	20.9
35-44	57	23.4	College	71	29.1
45-54	14	5.7	Bachelor's degree	93	38.1
55-64	6	2.5	Master's degree	20	8.2
65+	1	0.4	Advanced graduate work or PhD	7	2.9
			Other	1	0.4
Occupation	n	%	Occupation (continued)	n	%
Creative	7	2.9	Jobless	24	9.8
Administration	13	5.3	Retail	5	2.0
Academia / Research	5	2.0	Sales	5	2.0
Estate	4	1.6	Security	4	1.6
Finance	8	3.3	Service	13	5.3
Healthcare	14	5.7	Social	3	1.2
IT	27	11.1	Student	29	11.9
Law	6	2.5	Teaching	15	6.1
Literature / Writing	12	4.9	Technical	17	7.0
Management	20	8.2	Other	13	5.3

Table 3.1: Age distribution, highest level of education and profession of the 244 participants from the online survey.

3.3.3 Participants

The survey was active for one week, during which 272 submissions were made. After inspecting the control questions, we removed 28 submissions (10.3%) as some answers did not fulfill the desired requirements. Criteria were the time needed to complete the survey, the length of the answers and the answers to control questions. For example, two participants answered almost all open questions with the text “n/a”. Some other participants reported to own devices that were not smartphones and some other participants selected the wrong answer to the control question “Please select number four so that we know you are paying attention”. Thus, we were left with the responses of 244 survey participants (104 female, 42.6%). They were between 18 and 66 years old (average: 32 years). Most participants (191, 78.3%) were employed and had different professional backgrounds (e.g., design, healthcare, IT, etc.). The remaining participants were students (29, 11.9%) or unemployed (24, 9.8%). Most participants owned an Android device (139, 57%), 99 had an Apple device (40.6%) and 6 had a Nokia device (2.5%). Table 3.1 lists the age distributions, education and professions of our participants.

		SIM PIN	Lock Screen Authentication	Account
Used on Smartphone?	Yes	111 (45.5%)	165 (67.6%)	237 (97.1%)
	No	133 (54.5%)	79 (32.4%)	7 (2.9%)
Lockout Experience?	Yes	4 (3.6%)	56 (33.9%)	75 (31.6%)
	No	107 (96.4%)	109 (66.1%)	162 (68.4%)

Table 3.2: Overview of the number of participants that use a PIN for their SIM card, lock their lock screen (e.g., with PIN) or use web accounts on their smartphones. The table also shows the number of participants that have experienced lockouts.

3.3.4 Results

Table 3.2 gives an overview of the lockout frequencies for each lockout scenario. More detailed information about these lockouts is reported next.

SIM Card Lockout More than half of the participants (133, 54.5%) responded to have never used a PIN to protect their SIM card. Main reasons for this were the unawareness of the availability of a SIM PIN or the confusion with other concepts, such as the use of PIN codes for lock screen authentication (74, 55.6%). Being unconcerned about the security of their SIM card was mentioned by 45 participants (33.8%). Other reasons were, for example, the ownership of CDMA devices, inconvenience or fear of lockout.

While 111 participants (45.5%) reported to protect their SIM card with a PIN, only 4 of them (3.6%) had experiences with lockouts. The four participants provided different reasons for the lockouts. This includes forgetting, mistyping or confusing the PIN as well as the involvement of another person (e.g., mother who tries to access the phone). In order to unlock their SIM card, participants were required to enter the PUC, which not all participants were able to provide. Two participants had to contact the service provider to request a SIM card replacement/new PUC.

Lock Screen Lockout Altogether, 165 participants (67.6%) stated to use or have used lock screen authentication, such as PIN codes (106, 64.2%), graphical patterns (52, 31.5%), Touch ID (13, 7.9%) or face recognition (1, 0.6%). The numbers do not add up to 100% as some participants mentioned the use of multiple authentication schemes.

The experience of lock screen lockouts was reported by 56 (33.9%) participants. The reasons were mistypes due to hurry (14, 25%), inattentiveness (8, 14.3%) or being “*too drunk or otherwise intoxicated to correctly input the pattern*” (2, 3.6%). Other reasons were recent password changes (8, 14.3%), technical reasons (7, 12.5%), momentarily forgotten passwords (6, 10.7%) or the involvement of third parties, such as children playing with the phone (3, 5.4%).

In order to regain access to their devices after lockout, participants mentioned the use of PIN codes (e.g., as a fallback scheme for Touch ID; 4, 7.1%) or email-based resets (e.g., as

a fallback scheme for PIN codes; 2, 3.6%). However, the majority had to do nothing, but wait between 30 seconds and 15 minutes (48, 85.7%) before re-authentication was possible. In one particular case, a participant reported waiting times of several years and as a consequence bought a new phone (“*I couldn’t unlock the phone because my daughter reset the code and couldn’t remember it. Had to buy a new phone*”).

The use of waiting times was perceived ambiguously. While 21 out of 56 participants (37.5%) described them as easy and quick, other 21 out of 56 participants (37.5%) found them annoying, so that six out of 56 participants (10.7%) even removed lock screen authentication. However, fallback experiences were not the main reasons for removing lock screen security. Interestingly, some participants (13 out of 56, 23.2%) considered the lockout experience as a proof of concept of the security of a system (“*works and keeps other people out*”).

Account Lockout The majority of participants use special apps or mobile browsers to access online accounts on their smartphones (237, 97.1%). Seventy-five of them (31.6% reported lockout experiences.

The main reasons were the use of incorrect passwords due to password loss (50, 21.1%), password confusion (5, 2.1%) and typos (21, 8.9%). Other reasons were device constraints (6, 2.5%) or technical reasons (5, 2.1%). This included small keyboards, small screens or automatic logouts.

Most participants used traditional fallback schemes for password retrieval, such as email-links (63, 26.6%) or security questions (5, 2.1%). Other participants mentioned the use of password managers and written notes (5, 2.1%), while two participants noted that they usually postpone the password recovery until access to a computer is available (0.8%).

In general, participants did not notice major problems during account recovery, but some of them (45, 19%) criticized the usability of current schemes that often involved waiting times before recovery was possible. Nonetheless, the experience of being locked out of one’s account, similar to lock screen lockouts, was considered as a proof of security. This was mentioned by 23 participants (9.7%).

3.4 Individual Reports on Fallback Experiences

The results from the online survey suggest that fallback authentication does not happen frequently. However, for some users, recovery from lockouts seems to be more difficult than for others (e.g., one participant got rid of her phone). In order to gather more qualitative insights about these kinds of difficulties, we conducted semi-structured interviews (n=12) in July 2014. The design of the interviews, their evaluation and results are reported next.

3.4.1 Interview Design

The structure of the interviews was similar to the online survey, but used a between-groups design, meaning that each participant was questioned about only one lockout scenario (i.e., four interviewees per group). This was done to collect more details about the individual fallback experiences for each of the scenarios.

Interviewees were not participants from the online survey, but were recruited through mailing lists, bulletin boards and social media. It was a prerequisite to own a smartphone and to have experienced at least one of the three lockout scenarios in order to participate.

The interviews started with questions about the interviewee's demographics and smartphone usage. This was followed by the question whether interviewees have experienced [SIM card | lock screen | account] lockouts. Based on their response, interviewees were assigned to one of the three lockout scenarios for which they were encouraged to report more details about their last fallback experience. For example, they were asked where it happened, how it happened and what problems they encountered during fallback authentication.

The interviews were audio recorded, transcribed and later summarized by the interviewer for evaluation. Each interview lasted for about 30 minutes and interviewees received 5€ gift vouchers as incentive for their participation.

3.4.2 Participants

We interviewed 12 smartphone owners (6 female). The interviewees were between 17 and 55 years old (average: 27 years) and nine of them were students from different field of studies, such as media informatics, teaching, construction or law. Two interviewees were already employed in construction engineering or IT services, while one was a high school student.

Ten interviewees owned an iPhone (three with Touch ID). The remaining two had an Android device and Blackberry device, respectively. Interviewees reported to own their device for between 7 and 41 months (average: 20 months).

3.4.3 Results

SIM Card Lockout All interviewees stated to use a PIN to protect their SIM card, but only eight of them (66.7%) reported experiences with SIM card lockouts (Table 3.3).

The four interviewees that were assigned to this scenario provided additional details about how the lockout happened. Two of them explained that they owned multiple SIM cards with different PINs that they mixed up during authentication. The remaining ones stated that their smartphone battery died unexpectedly so that they had to enter the SIM PIN. However, both of them could not recall the exact PIN and tried multiple combinations until their devices got locked completely.

		SIM PIN	Lock Screen Authentication	Account
Used on Smartphone?	Yes	12 (100%)	11 (91.7%)	12 (100%)
	No	0 (0%)	1 (8.3%)	0 (0%)
Lockout Experience?	Yes	8 (66.7%)	9 (81.8%)	11 (91.7%)
	No	4 (33.3%)	2 (18.2%)	1 (8.3%)

Table 3.3: Overview of the number of interviewees that use a PIN for their SIM card, lock their lock screen (e.g., with PIN) or use web accounts on their smartphones. The table also shows the number of interviewees that have experienced lockouts.

None of them was able to unlock the SIM card without additional help. For example, one interviewee was locked out during a vacation in Spain. The interviewee was stuck in an authentication deadlock as the required PUC was not at hand and the device was locked so that no calls could be made. The interviewee was lucky to be able to borrow the phone of a friend to call a family member who looked up the PUC. However, the interviewee complained about the long waiting time (about an hour). This was also mentioned by other interviewees, who found it annoying to wait up to several days until the service provider unlocked their SIM card.

Lock Screen Lockout Eleven interviewees (91.7%) claimed to protect or have protected their lock screen (Table 3.3). The protective measures mentioned were: PIN Code (7; 63.6%), Touch ID (3; 27.3%); Android Pattern (1; 9%) and alphanumeric password (1; 9%). Lockouts were experienced by nine interviewees (81.8%).

Two interviewees assigned to this scenario were Touch ID users who reported failed finger-print recognition as reasons for lockouts. The other two in this group used PIN codes for authentication. One of them was locked out due to being in a hurry, while the other one reported that the lockouts were caused by friends who tried to play a prank. This happened multiple times and led to waiting periods that ranged from a few minutes to several days. In the latter case, the corresponding interviewee was too impatient to wait until the timeout period was over and took the phone to a store to get the lock removed. In turn, the Touch ID users did not have to wait, but had to enter a PIN code instead. Both of them noted that recalling the PIN code was not difficult as they reused a code that they had already used on an older phone. Both Touch ID users mentioned higher lockout frequencies than the other interviewees. They experienced lockouts on a monthly or even weekly basis.

Account Lockout All interviewees noted that they use their smartphones to access online accounts and 11 of them (91.7%) had experiences with account lockouts (Table 3.3).

Three of the four interviewees in this group gave forgetting their passwords as lockout reasons. For example, one interviewee explained that she usually saves her passwords to be logged in automatically, making it difficult to recall them when she gets logged out (e.g.,

when friends borrow the phone to access their online accounts). Nonetheless, all three described the fallback procedure as simple.

One interviewee found her last fallback experience very difficult. She was on vacation when she tried to access her Facebook account using the corresponding app. Though she was sure that she entered the correct password, she was not able to log in. The same attempt failed when she tried it on her boyfriend's phone. Furthermore, the app did not display any details about the problem. The problem was caused due to the unavailability of a desktop computer. Later, when she had access to one, she found out that she had to complete additional authentication tasks by identifying persons on photos to login.

3.5 Discussion

The online survey gave a general overview of fallback authentication in the wild and was complemented by interviews that provided further details about individual lockout experiences. The results suggest that lockouts are experienced infrequently, but regaining access to the device or account becomes difficult when certain circumstances apply, for example, when smartphone owners are out of their usual context (e.g., vacation).

This does not mean that current solutions need to be replaced, but leave room for improvements. For example, waiting times can be annoying so that the use of fallback schemes may be more convenient to the user. These schemes exist, but are seldom shown to the user or sometimes not usable in lockout situations due to their dependency from third parties. The following discussion summarizes these problems and proposes ideas that should be considered in the future design of fallback schemes.

3.5.1 Lack of Information

Most participants had no difficulties during fallback authentication and knew what they had to do. However, some participants were not able to log in as they were caught in an authentication deadlock (e.g., the provided fallback option did not work out). For example, one participant forgot her PIN code that she entered incorrectly multiple times in a row. This resulted in a waiting period. Since she had forgotten her PIN code, she ended up with increasing waiting times.

Therefore, users should be informed by the system about other alternatives when they repeatedly fail a particular fallback scheme. For example, users that reach unacceptable waiting times of multiple years should be advised that there is also the possibility of email-based password resets and, as a last resort, the option to reset their device to factory settings.

3.5.2 Feeling Secure, but Annoyed

The availability and experience of fallback authentication was considered by participants as a proof that their device is secure against unauthorized access. However, lockouts were also often described as annoying, mostly due to idle time periods in which users could do nothing but wait.

Previous work has shown that waiting is often associated with boredom [31] and has a negative impact on mood, customer satisfaction as well as perceived duration [25]. To improve the subjective assessment of time, Chebat and Filiatrault suggest the creation of interesting tasks [25]. Though this suggestion was made in the context of customer service, it seems worth exploring for fallback authentication as well. Instead of waiting times, alternative fallback solutions could engage users more actively in the authentication process by creating small security tasks, such as drawing scribbles (e.g., [134]) or asking users about their recent phone activities (e.g., [37]). Though these kinds of tasks take longer than the actual waiting times, the perceived duration may be shorter.

3.5.3 Third-Party Dependencies

Most participants who encountered difficulties during fallback authentication were not able to solve the problems by themselves. Instead, they had to rely on other persons that, for example, looked up information (e.g., PUC) or that provided them with technical equipment (e.g., phone or computer) that was not at hand due to special circumstances (e.g., vacation).

Most common fallback solutions do not take these problems into account and some of them do not even adapt to the mobile environment, meaning that they can only be used on non-mobile devices. Thus, it seems advisable to complement existing solutions with alternative options that can be used immediately (to reduce waiting times) and that are independent from information that is hard to obtain when the device is blocked (e.g., limited Internet access).

3.5.4 Third Parties

Most lockouts were self-inflicted, but participants also reported on lockouts that were caused by close persons (e.g., friends, children). Some of them caused lockouts unintentionally, while others did this on purpose.

This suggests that alternative fallback schemes must be at least as secure as the primary authentication scheme to prevent the latter type of adversary (i.e., adversaries that try to authenticate as the legitimate user) from circumventing other authentication schemes to gain access to the device.

3.6 Lessons Learned

In this chapter we have presented three scenarios in which fallback authentication on smartphones may take place and evaluated their actual occurrences in the wild through an online survey and complementary interviews. Since only few smartphone owners seem to have difficulties during fallback authentication, the reported problems are not representative of the general population. Nonetheless, there are users that are unable to regain access with currently available solutions when experiencing lockouts (i.e., authentication deadlocks). The main goal of this thesis is to support this target group and to explore alternative fallback schemes whose design requirements are based on the reported problems. These requirements can be summarized as follows:

- **Independence.** Alternative fallback schemes should be independent of additional components (e.g., persons, technical equipment or tokens). This is important to ensure the applicability of the fallback schemes in any kind of situation.
- **Immediacy.** Alternative fallback schemes should be immediate. For example, users should not have to wait multiple days before they can retry authentication.
- **User engagement.** Alternative fallback schemes should require the active engagement of users in the overall fallback process to be less likely to cause annoyance.
- **Security.** Alternative fallback schemes should be at least as secure as the primary authentication as the security of a system is defined by its weakest link, meaning that each additional alternative represents a potential point of attack for adversaries.

4

A Framework for Fallback Authentication

The objective of this chapter is to motivate our design decisions and our evaluation strategies. It starts with a formal definition of fallback authentication that elaborates on the similarities and differences of primary authentication to fallback authentication (Section 4.1).

Based on this, we derive a design space that drives the projects presented in this work. Each project is briefly summarized to illustrate what is to be expected in the remainder of this thesis (Section 4.2). This is followed by an overview of our evaluation strategies that takes into account usability factors (e.g., memorability) and security factors (e.g., human threats) to identify the best trade-off between the two by using the accuracy metric (Section 4.3).

The chapter concludes with a summary of the key aspects of the presented framework for fallback authentication (Section 4.4).

4.1 Fallback Authentication

So far, we have described fallback authentication as an alternative authentication scheme that is used when the primary scheme fails. A more formal definition is developed next.

4.1.1 The Authentication Chain

According to Renaud et al., authentication is a three-stage process that consists of enrollment, authentication and replacement [143]. Enrollment refers to the registration of the user with the system where a secret key for authentication is defined (e.g., password). This key is then used during authentication to get access, but needs replacement when users forget it.

Fallback authentication is part of the third stage when users lose their secret key. Since a new key cannot be issued without verifying the user's identity, replacement (i.e., fallback authentication) also encompasses the three stages of enrollment, authentication and replacement. In this context, enrollment refers to the process where users share an alternative secret with the system (e.g., by answering security questions) that is then used as an alternative means of authentication. However, if this alternative secret is forgotten as well, a replacement for the replacement is needed.

Based on this consideration, the authentication process is a chain of multiple authentication schemes. It starts with a primary scheme and is followed by several alternatives that are used successively when replacement is needed (Figure 4.1). A practical implementation is the authentication chain found on Android devices. It starts with Face Unlock (if set up) and continues with a graphical pattern, email-based reset and factory reset as fallback options.

4.1.2 Classification of Fallback Authentication

Since fallback authentication is only a special occurrence of authentication in general, both of them can use the same classification, meaning that one can distinguish between knowledge-based schemes (i.e., something you know), biometric schemes (i.e., something you are) and token-based schemes (i.e., something you have) [187]. However, this raises the question of the differences between the two, if the same classification is applied.

An important factor in this regard is their scenario of use [86]. Primary authentication is the main authentication scheme of a system and needed more frequently than fallback options. The latter is only required when the former fails, which, in the best case, should never happen. Transferring this view to the authentication chain implies that schemes located at the end of the chain are needed less frequently than those at the beginning.

As a consequence, different usability and security requirements must be considered in the design of primary schemes and fallback schemes [85]. Factors that are important for the former may weigh less in the design of the latter and the other way around. Kainda et

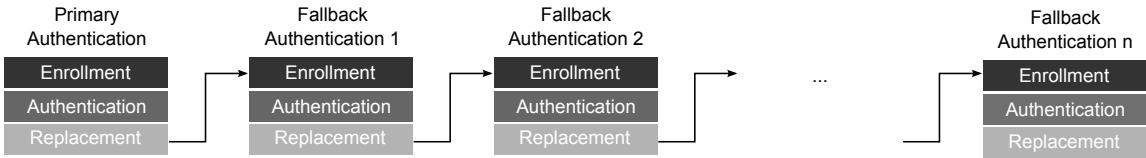


Figure 4.1: Fallback authentication chain based on the three-stage authentication procedure described by Renaud et al. [143].

al. identified several factors that are commonly used for the evaluation of authentication schemes [104]. We will summarize them in the following and discuss them with respect to their importance to primary authentication and fallback authentication.

Effectiveness An authentication system is effective when it is able to distinguish between legitimate users and illegitimate ones. This factor is essential for all authentication schemes. However, in the best case, all legitimate and illegitimate users are identified correctly by the primary scheme to render the need for fallback schemes obsolete as the latter is only required when the primary scheme becomes ineffective.

Accuracy The accuracy of an authentication system is closely related to its effectiveness. It depicts the success rate of the system in distinguishing between legitimate and illegitimate users and thus similar to the effectiveness, is important for all authentication schemes. More details about the accuracy measure will be discussed later (Section 4.3.3).

Efficiency The effort or time that users need to authenticate is referred to as efficiency. This factor is critical for primary authentication due to its frequent use (e.g., lock screen authentication), but has a lower priority in the context of fallback authentication. However, efficiency-related requirements also depend on the scenario of use, meaning that schemes located at the beginning of the fallback chain are more time-sensitive than those at the end.

Satisfaction User satisfaction is a crucial factor to prevent users from disabling security measures or exhibiting insecure behaviors. Though satisfaction, to some extent, is important for all authentication schemes, it should have a higher impact on the design of primary schemes due their frequency of use. For example, even though most fallback experiences are annoying, they are, in the most cases, not the main reason for removing security measures (see Chapter 3).

Memorability The ease with which users can recall their passwords during authentication relates to the memorability of an authentication scheme. In particular, knowledge-based schemes are known for their memorability issues due to the number of passwords that users have to cope with (Section 2.1). While memorability is indisputably important for primary authentication, it is even more important for fallback schemes as they are needed less frequently so that recall becomes more challenging.

Learnability The ease with which users can learn how to use an authentication scheme is called learnability. It is the first hurdle that users encounter when they authenticate/enroll for the first time. Although learnability is important for primary as well as fallback authentication, the reasons for this are different. While primary schemes need good learnability to avoid high dropout rates, fallback schemes need it due to the lack of training that users have with the corresponding scheme. In addition to this, users may not be aware of the importance of fallback authentication (e.g., they may assume that they will never need it) and so learnability is essential to encourage them to set up these schemes.

Motivation Users have different motivations to protect their data which, in turn, impacts their password selection (Section 2.1). For example, users may use longer passwords for sensitive accounts and shorter ones for unimportant accounts. The different motivations are relevant for the design of primary schemes, but also for fallback schemes as they can impact the time that users are willing to spend for enrollment.

Social Context Humans are often surrounded by other humans that interact and communicate with each other (e.g., family, friends or strangers). But humans can also turn into potential adversaries that take advantage of shared secrets to compromise systems. These adversaries exist both for primary and fallback schemes, but the threats that are of concern may differ.

In summary, many factors are important for the design of primary schemes as well as fallback schemes. They influence different stages of our research: Learnability and motivation inspire various of our designs to enable authentication without prior enrollment by the user; memorability and efficiency are key factors of our evaluation strategy, while the factor social context plays a particular role in the definition of potential threats for our security analysis. In turn, we find efficiency and satisfaction less relevant due to the infrequent occurrence of fallback authentication.

4.1.3 Definition of Fallback Authentication

Based on the preceding comparison between primary authentication and fallback authentication, we derive the following definition for fallback authentication:

Definition.

Fallback authentication is part of an authentication chain that consists of $1 \dots n$ authentication schemes, with $n \in \mathbb{N}$. Each authentication scheme in the chain involves a three-stage process: enrollment, authentication and replacement. An authentication scheme is called fallback scheme if it is preceded by another authentication scheme in the chain. Otherwise it is referred to as the primary scheme.

Fallback schemes should further fulfill the following requirements:

Effectiveness and Accuracy

Fallback schemes must be at least as secure as the authentication schemes that precede them in the chain to ensure security.

Efficiency

Fallback schemes are less time-sensitive than the schemes that precede them, but more time-sensitive than the ones that follow them. This is done to reflect their frequency of use and to define their order of issue.

Learnability

Fallback schemes must be usable without prior training, meaning that users should immediately know how to authenticate when they encounter the scheme for the first time.

Memorability

Fallback authentication must not rely on information that is learned by heart (e.g., through repetition), but should take advantage of recent memories or memories that are less likely to be forgotten over time (e.g., episodic memories).

Satisfaction

Fallback schemes can be, but do not have to be satisfying due to the infrequent occurrence of lockouts.

Motivation

The enrollment effort for fallback schemes must be kept to a minimum to motivate users to set them up.

4.2 Exploring the Design Space

The design space for alternative fallback authentication schemes can be defined along four dimensions: type of replacement, type of enrollment, type of challenge and type of response (Table 4.1). These dimensions are based on the three authentication stages and are discussed in more detail in the remainder of this section.

4.2.1 Type of Replacement

The type of replacement (i.e., the type of fallback scheme) can be categorized as biometric, knowledge-based and token-based. Since the focus of this thesis lies on the use of autobiographical memories, we will only cover knowledge-based fallback schemes and behavioral fallback schemes within the design space.

4.2.2 Type of Enrollment

Enrollment is required to define a shared secret between the user and the corresponding system. It can occur explicitly (e.g., [34]) or implicitly (e.g., [40]). While the former requires active involvement by the user, the latter happens in the background, for example, during the user's interaction with the device. With respect to the secret shared, explicit enrollment deals with static information that only changes when initiated by the user, while implicit enrollment handles dynamic information that is updated automatically with each interaction. The remainder of this thesis will refer to them as static enrollment and dynamic enrollment.

4.2.3 Type of Challenge and Type of Response

Authentication is based on a challenge-response model where the system confronts the user with a challenge that the user has to respond to. Based on the distinction by Just [101], we classify challenges and responses as fixed, controlled and open.

4.2.4 Overview of Example Implementations

Based on the given design space, there are $2 \times 3 \times 3 = 18$ possible combinations for the three dimensions of enrollment (with two levels), challenge (with three levels) and response (with three levels). For each of the possible combinations, numerous implementation options exist.

Overall, we explore twelve combinations for the design of fallback schemes. These schemes cover static and dynamic enrollments, but focus on the latter due to motivational factors (see Section 4.1) and the known problems of static enrollments (see Section 2.3). The type of

		Challenge			Enrollment
		open	controlled	fixed	
Response	open	❶ -	❶ -	❶, ❷ -	static dynamic
	controlled	❶ -	❶ -	❶ ❷	static dynamic
		-	-	❷	static dynamic
	fixed	-	-	❸, ❹, ❺	dynamic

- ❶ Location-Based Questions ❷ Icon Arrangements
 ❷ 2-Finger Sketches ❸ Installed Apps
 ❸ Smartphone Activities

Table 4.1: Design space for alternative fallback schemes based on a challenge-response model. The numbers refer to the approaches covered in this work and their classification within the design space. A detailed description to the corresponding numbers can be found in section 4.2.4. Cells with a vertical score represent combinations that are not viable, while empty cells are combinations that are not covered in this thesis.

challenge and the type of response was chosen based on their suitability for the corresponding scheme. For example, we focused on fixed challenges for dynamic enrollments as they happen without the active involvement of the user, rendering the possibility to define own challenges obsolete. We further favored controlled responses over open responses as the latter is only a more general version of the former. An overview of the different fallback schemes covered in the scope of this thesis and their classification within the design space is given in Table 4.1. The remainder of this section will provide a brief description for each fallback scheme covered in this work. The schemes are sorted by their type of enrollment.

Static Enrollment

❶ Location-Based Questions Personal experiences are remembered with a strong spatio-temporal context (Section 2.6). We exploit this characteristic for the design of location-based questions (e.g., “*Where did your first kiss take place?*”). In comparison to traditional security questions, the answers are not provided as text, but as locations on a map. We evaluate how accurate these locations are recalled by users and how easy they are guessed by different types of adversaries.

❷ 2-Finger Sketches The use of biometric features from drawn sketches has produced promising results in the context of primary authentication (e.g., [40]). Therefore, we explore if sketch-based authentication can also be exploited for fallback authentication. We designed a fallback scheme that requires users to reproduce predefined sketches with two fingers simultaneously. This was done to positively influence the recall of stroke order which is an important feature for biometric authentication.

Dynamic Enrollment

③ Smartphone Activities Many interactions with the smartphone are part of autobiographical memory, such as a call with one's significant other or a photo taken during a vacation. In this approach, we examine the suitability of different smartphone activities (e.g., receiving a call or taking a photo) for the design of alternative fallback schemes.

④ Icon Arrangements Smartphone owners start a substantial amount of apps from home screens (i.e., the mobile counterparts to desktops on computers) [73]. Since interaction with these screens happens regularly, it is assumed that users implicitly learn the arrangement of their apps [71]. This approach evaluates the accuracy with which users can reproduce their home screen layout as well as the suitability of icon arrangements for fallback authentication.

⑤ Installed Apps Previous work has shown that app usage is closely related to a person's personality [30]. For example, introverted people exhibit other smartphone usage patterns than extroverted people. This approach explores the potential of installed apps for the design of fallback schemes where users authenticate by deciding whether certain apps are or are not installed on their device.

4.3 Research Strategy

For each of the example implementations, our research strategy is threefold: (1) we demonstrate how different types of autobiographical memories can be leveraged in fallback schemes; (2) we evaluate the usability (i.e., memorability) and security of these memories for fallback authentication; and (3) we identify the best trade-off between the two factors (i.e., usability and security) and discuss their implications on the design of fallback schemes.

Since a brief overview of the example implementations has already been given in the previous section, the remainder of this chapter will focus on the evaluation methodology in this work and the metrics used.

4.3.1 Usability Evaluation Methodology

Memorability is one of the main factors that we will consider for usability evaluation, since it was shown to be one of the major issues of current knowledge-based solutions for fallback authentication (see Section 2.3). Memorability can be measured by the ratio between the number of correct responses and the number of all responses given by users (i.e., success rate [104]). However, this ratio is influenced by different variables that are explained next.

Time Passed The time that passes between enrollment and fallback authentication is likely to influence memorability as the recall of information is prone to deteriorate over time. Thus, it is important to conduct short-term and long-term memorability tests. This is a common approach in the domain of fallback authentication (see Section 2.3).

Number of Challenges The number of challenges that users have to respond to has an impact on factors like memorability, efficiency and security. For example, the more challenges are required, the higher the burden is on memorability and the longer the authentication lasts. Nonetheless, the use of multiple challenges is well-established to prevent random adversaries from authenticating successfully by chance.

Number of Attempts Since the increasing number of challenges can cause users to make more errors, it is advisable to give them multiple attempts to provide the correct responses in case they fail. However, it is important to limit the number of attempts as it increases the risk of potential adversaries to guess the right answers.

In summary, memorability is a key factor for the usability evaluation of fallback schemes. In the scope of this thesis, we will evaluate memorability at different points in time after enrollment to test the short-term memorability and long-term memorability of the corresponding fallback schemes (e.g., immediately after enrollment and six months later). Memorability will be measured by the success rate of users to recall their responses, but will also take into account parameters like the number of challenges and the number of allowed attempts that are crucial to decide whether an authentication is considered as successful or not. More details on this will be provided later (Section 4.3.3).

4.3.2 Security Evaluation Methodology

The security of an authentication scheme can only be analyzed when potential threats are identified and their prevalence is estimated. This is done to decide against which threats an authentication scheme should protect [86]. Our threat models focus on threats that are based on the ones identified by Just and Aspinall [103]: insider attacks, focused attacks and blind guesses. We classify these threats according to the relationship between the adversary and the victim as this factor was shown to be important by past work (e.g., [128, 170]) and was also observed during the survey and interviews (see Chapter 3). With respect to other types of threats that are less prevalent (e.g., shoulder surfing), we will discuss them individually in the corresponding chapters in case these threats apply.

Insider Attacks The first threat model assumes a close adversary (e.g., significant other, best friend) who is in possession of the victim's device. In order to access it, the adversary tries to guess the victim's password, but fails so that the device gets locked and fallback authentication is required. Since the adversary has advanced knowledge about the victim, guessing the right answers is not plain luck anymore.

Focused Attack The second threat model assumes a stranger as adversary who has gotten hold of the victim’s device (e.g., by stealing it). The adversary wants to get access to the device, but gets locked out completely and thus tries to exploit the fallback scheme. The adversary further researches the responses to the challenges (e.g., from social networks) so that guessing the right answers is above chance.

Blind Guesses The third threat model also assumes a stranger as adversary who has gained physical access to the victim’s device (e.g., by finding it on the street). The adversary tries to access the device, but gets locked out completely so that fallback authentication is required. Since the victim is unknown, the stranger guesses the answers arbitrarily.

In summary, there are different kinds of threats in the context of mobile fallback authentication that are motivated by the adversary’s relationship to the victim. In the scope of this thesis, all three threats will be considered (where appropriate), but with a focus on insider attacks as they are one of the worst-case scenarios. This means that we will evaluate the fallback schemes not only from a theoretical point of view, but also with human adversaries that have a close relationship to the user.

4.3.3 Accuracy Metric

In order to analyze the best trade-off between usability and security of a fallback scheme, we will take advantage of the accuracy metric which is a good indicator for how well a system works [121]. This metric is used in various domains, such as medical diagnoses, physics, chemistry or security research.

In general, accuracy is represented by the ratio of the number of correct decisions to the number of all decisions made by a system [121]. In the context of authentication, correct decisions are made when legitimate users are authenticated successfully, while adversaries are rejected rightfully. These decisions are also referred to as true positives (TP) and true negatives (TN), respectively. In turn, incorrect decisions are referred to as false negatives (FN) and false positives (FP), meaning that legitimate users are rejected, while adversaries are accepted. In other words, the accuracy for an authentication scheme is the ratio of correct authentication decisions to all authentication decisions made and can be described by the following formula:

$$Accuracy = \frac{\sum TP}{\sum TP + \sum FN + \sum TN + \sum FP} + \frac{\sum TN}{\sum TP + \sum FN + \sum TN + \sum FP}$$

The formula returns a value between 0 and 1 (0 and 100 in percentage values). In the best case, the accuracy is 1 (or 100%), meaning that all legitimate users and adversaries are identified as such.

Accuracy values must be interpreted carefully. For example, two fallback schemes can yield the same accuracy values, but still perform differently: while the incorrect decisions of one

scheme may consist of FN only, the incorrect decisions of the other may all be FP. This leads to different implications for the usability and security of a system. Therefore, it is important to complement accuracy values with two types of success rates that are referred to as sensitivity and specificity [121]:

$$\text{Sensitivity} = \frac{\sum TP}{\sum TP + \sum FN} \quad \text{Specificity} = \frac{\sum TN}{\sum TN + \sum FP}$$

Sensitivity refers to the success rate of users and is depicted by the ratio of successful authentication attempts to all attempts made by them. In turn, specificity is the failure rate of adversaries, represented by the ratio of successfully rejected attacks to all attacks made. Thus, a high sensitivity implies good usability, while a high specificity implies good security.

So far, we have mentioned successful and unsuccessful authentication attempts. However, it is still unclear when an authentication attempt is considered as successful or unsuccessful. For this, a so-called decision threshold is needed that is influenced by various implicit variables [121]. In the context of challenge-response systems for authentication, these variables are the number of challenges that needs to be answered correctly as well as the number of attempts allowed to respond to each challenge (see Section 4.3.1). Since these parameters influence the number of FP and FN, they will also have an impact on the accuracy values.

In summary, the main motivation of the accuracy metric in the scope of this work is to quantify the trade-off between usability and accuracy. This includes the identification of the best parameters (i.e., number of challenges, number of attempts) and their impact on the success rates of users and adversaries. This is important to discuss design implications for possible real-world deployments.

4.4 Chapter Summary

In this chapter we have provided a theoretical framework for fallback authentication to motivate the design and evaluation of alternative fallback schemes that will be presented in this work. The content of this chapter can be summarized as follows:

- **Definition of Fallback Authentication.** Fallback authentication is a chain of alternative authentication schemes that are issued successively if the primary scheme fails. Although any scheme for primary authentication could be used for fallback authentication, their scenario of use imposes different requirements on their design.
- **Differences to Primary Authentication.** The main difference between primary authentication and fallback authentication is their frequency of use so that certain usability and security factors become more/less important in the design of one or the other. For example, while efficiency is a crucial aspect for primary authentication, it is less critical in the context of fallback authentication. Instead, the factor memorability should have a higher priority for the latter.

-
- **Design Space for Fallback Authentication.** The design space for fallback authentication can be defined along four dimensions: type of replacement, type of enrollment, type of challenge and type of response. In the scope of this thesis, we focus on replacement schemes that have a relation to autobiographical memory. The different types of memories are explored in various approaches that are either based on static or dynamic enrollments.
 - **Evaluation Strategy.** Since most fallback schemes were shown to be lacking in memorability, the usability evaluation in this work focuses on memorability-related factors (i.e., number of recalled items, number of challenges and number of allowed attempts). In terms of security, different types of threats are considered that take into account the social context of the user (e.g., insider attacks). Last, but not least we use accuracy calculations as a metrics to identify the best trade-off between usability and security to discuss the potential of different types of memories for fallback authentication.

5

Fallback Authentication Based on Static Enrollments

The use of static enrollments for fallback authentication has the advantage that users are actively involved in the definition of the secrets that they share with the system they register for. This way, they get to know the authentication scheme before the actual lockout happens and thus are not unprepared when fallback authentication takes place.

However, static enrollments also have their shortcomings, such as requiring users to spend a certain amount of time to provide the needed information. More seriously than that, they often cause memorability issues when their design does not take into account that a long time may pass after enrollment before the provided information is needed again.

In this chapter, we evaluate two approaches that use declarative and non-declarative knowledge about autobiographical memories for fallback authentication. While one of them focuses on the skill to draw (Section 5.1), the other takes advantage of the strong spatio-temporal context of episodic memories to create location-based questions (Section 5.2). These types of memories are, even after longer periods of time, easier to recall than personal facts that are often used for traditional security questions (see Section 2.6).

We found that drawing skills (in our implementation) do not work for authentication. The presented fallback scheme is a prime example that memorability is not the only usability factor that is important for fallback authentication; user satisfaction is also essential for an authentication scheme to work. With respect to location-based questions, the results are very promising and we see a high potential for their deployment in a real-world setting (Section 5.3).

Personal contribution statement: The content of this chapter is based on two bachelor theses by Lara Hirschbeck [89] and Michael Richter [145]. Both theses were supervised by the author. Part of this work was published at the Symposium on Usable Privacy and Security (SOUPS 2015), with co-authors Alexander De Luca, Michael Richter, Matthew Smith and Heinrich Hussmann [76]. A detailed overview of task responsibilities can be found in the disclaimer.

5.1 Sketch-based Fallback Authentication

The ability to sketch is a procedural skill that forms part of autobiographical memory (see Section 2.6). It has often been used in the domain of biometric authentication due to the unique features that drawings convey. Popular examples are draw-a-secret [98], passdoodles [66] and scribble-a-secret [134]. All of them allow users to create free-hand sketches on a grid or canvas and assume better memorability based on the fact that pictures are better remembered than words [129]. In this section, we introduce a sketch-based authentication scheme for fallback authentication. We present its design, implementation and evaluation to report the key insights from the latter.

In general, sketch-based authentication schemes are very similar: Users create an authentication template by repeatedly drawing their passwords during enrollment. This template is then used during authentication to compare it with the authentication attempt. The comparison returns a value that describes the similarity between the two and if it exceeds a predefined threshold, the attempt is considered as successful. Otherwise the attempt is rejected.

The comparison itself can be done with different methods: Some approaches are inspired by signature verification and use visual matching for comparison (e.g., [4, 134, 171]). Others rely on implicit features that are collected during drawing and include, for example, the speed with which the sketch is made or the pressure that is used (e.g., [40, 67, 179]).

5.1.1 Approach

Since the results from previous research on sketch-based authentication were promising (e.g., [40]), we were keen on exploring its use in the context of fallback authentication on smartphones. We propose an approach that authenticates users by the way they draw a predefined sketch template on a smartphone touch screen using two of their fingers simultaneously (see below for further details).

The use of predefined sketch templates (instead of user-chosen ones) was inspired by the work of Alzubi et al. who provided users with drawing instructions during password selection (e.g., “*draw three connected wheels of different sizes*”) [4]. However, we do not use textual guidelines, but show the actual sketch that users have to draw for authentication. This was done to minimize potential memorability burdens [169].

Furthermore, drawing with two fingers simultaneously was motivated by the problems that Goldberg et al. identified [66]. Although many methods for sketch comparison use stroke order as distinction feature, users often have difficulties to recall the exact order. With the use of two fingers at the same time, we assume to positively influence stroke order due to physical restrictions. In addition to this, the use of two fingers allows us to explore additional features, such as the distance between two fingers, to distinguish between users.

T1	T2	T3	T4	T5	T6	T7	T8	T9

Table 5.1: Overview of the nine sketch templates (T1-T9) that were created during brainstorming sessions.

5.1.2 Design of Predefined Sketches

A brainstorming with four participants (all male) was conducted to collect ideas for different sketch templates. Participants were recruited through mailing lists, social media or personal communication. They were between 19 and 23 years old (average: 22 years) and had a background in media informatics. All of them owned a smartphone and were familiar with the capabilities and limitations of touch input on mobile devices. The brainstorming lasted about 30 minutes and participants received 5€ gift vouchers as incentive.

Participants were invited to our lab. After a brief overview of our approach, we asked them to come up with ideas for sketches that they think can be drawn on smartphone touch screens using two fingers simultaneously. We collected all ideas on a whiteboard for later discussion.

The participants' ideas were inspired by a variety of topics: A substantial amount of suggestions was based on different character sets, such as hieroglyphs, Chinese characters or Latin letters (e.g., Table 5.1, T9). Other ideas took into account finger movements and included sketches that required both fingers to move (either in the same or opposite direction) as well as sketches that required the fixation of one finger, while the other had to be moved (e.g., Table 5.1, T3 or T5). Symmetry and simplicity were also important factors: participants often proposed sketches that were composed of simple forms, such as lines, circles or triangles (e.g., Table 5.1, T8).

Table 5.1 overviews the final sketch templates that we used for evaluation. They were all based on the suggestions made by participants during the brainstorming.

5.1.3 Prototype

The study prototype was implemented for smartphones using Android 4.0 (Ice Cream Sandwich) or higher, but was optimized for our study device (i.e., Samsung Galaxy S3). Using the same device for all participants had the advantage that the collected data could also be used to simulate attacks for our security analysis. More details about this will be provided later (see Section 5.1.4).

The overall structure of the study application was simple (Figure 5.1). The opening screen explained the study task and as soon as participants pressed the OK button, they were successively shown the nine sketch templates that appeared in random order. For each template, participants were asked to draw them on a white canvas ten times in a row. In case they were

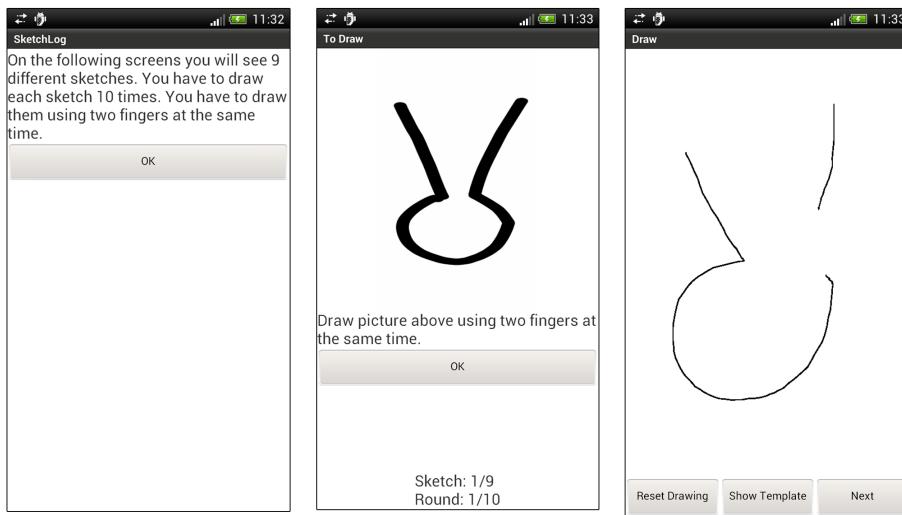


Figure 5.1: Screenshots of the study prototype: Opening screen (left), example of sketch template with instructions (center) and example of a sketch drawn by a user (right).

not satisfied with the result, participants had the option to reset their sketch to draw it again. The same application was used to collect the data for enrollment as well as authentication.

In order to detect the participants' interaction with the touch screen, we registered different touch event properties. For example, we identified whether the users' finger touched the screen, was lifted from the screen or was moved on the screen. Due to multi-touch support, this was done with both fingers simultaneously. Furthermore, the first finger to touch the screen was recognized as primary touch, while the other was considered as secondary touch.

Data collection was done every 25 milliseconds and all relevant information was stored in a CSV-file on the device. Each entry in the file consisted of the following parameters: event time, xy-coordinates of primary touch and xy-coordinates of secondary touch. In addition, we also stored the drawn image for later visual inspection.

5.1.4 Threat Model

Since predefined sketch templates reveal the secrets to be used for authentication, potential adversaries (e.g., strangers) who are in possession of the victim's device can try to attack the fallback scheme by simply drawing the templates (i.e., blind guesses; see Section 4.3.2).

In order to evaluate whether these kinds of adversaries are likely to succeed, we simulated attacks using the authentication attempts from the study. How this was done will be described later in the next section.

5.1.5 User Study

The user study consisted of two sessions: The first session was used for enrollment, while the second session (i.e., four weeks after the first session) simulated fallback authentication. A detailed description of the study is provided next.

Study Design and Study Procedure

Participants were recruited over bulletin boards, social media and personal communication. All participants completed the same tasks for the study. The independent variable was the *sketch template* (nine levels) to be drawn for enrollment/authentication. The order in which the templates appeared was randomized to minimize learning effects. The study design was the same for both sessions.

Session I (Enrollment) Participants were invited to our lab for the first session. At the beginning of the study, we gave them a brief introduction to the general procedure and explained the study task. Once they were ready, they started the study application and drew each of the nine sketch templates ten times in a row. There was a break between each template to hand participants a brief questionnaire about how easy it was to draw the corresponding sketch. The first session was concluded with another questionnaire that asked for demographic information and the participant's three most favorite templates.

Session II (Fallback Authentication) The second session was conducted four weeks after the first session to simulate fallback authentication. The procedure was the same as for the first session. The only difference was that the data from the first session was used to create the authentication template (i.e., enrollment), while the data from the second session was considered as actual authentication attempts and compared to the data from the first session.

Each session lasted about 60 minutes and participants received 20€ gift vouchers for their participation (10€ per session). Participants were required to show up for both sessions in order to receive their compensation.

Participants

General Demographics Altogether, 21 participants (4 female) took part in the experiment. They were between 11 and 58 years old (average: 30 years). Ten participants were students with a background in media informatics. Eight participants were already employed and worked as research assistants, physicians, electricians or insurance salesmen. One participant was a homemaker, while two others were high school students.

Drawing Skills When asked to rate their drawing skills on touch screen devices (using their fingers) on a 5-point Likert scale (1=very bad; 5=very good), most of them (15 participants) stated to have mediocre drawing skills (equals three on the Likert scale). Two participants were good in drawing on touch screen devices, while four participants stated to have bad or very bad drawing skills.

Combination	Event Time	X₁	Y₁	X₂	Y₂	Distance
1		✓	✓			
2				✓	✓	
3						✓
4		✓	✓	✓	✓	
5		✓	✓			✓
6				✓	✓	✓
7		✓	✓	✓	✓	✓
8	✓	✓	✓	✓	✓	
9	✓	✓	✓			✓
10	✓			✓	✓	✓
11	✓	✓	✓	✓	✓	✓

Parameter	Description
Event Time	Timestamp of touch event.
X₁	X-coordinate of first finger to touch the screen.
Y₁	Y-coordinate of first finger to touch the screen.
X₂	X-coordinate of second finger to touch the screen.
Y₂	Y-coordinate of second finger to touch the screen.
Distance	Distance between xy-coordinates of first and second finger.

Table 5.2: Overview of the parameter combinations used for the DTW analyses (top) and the description of the corresponding parameters (bottom).

Data Analysis

In order to analyze the data from the two sessions, dynamic time warping (DTW) was used. Although there were other options available (e.g., machine learning algorithms), DTW was shown to be a good option by previous work (e.g., [40]).

DTW originates from speech recognition, but has found usage in various domains related to pattern recognition [10]. Its general idea is to compare two time sequences to each other by calculating a so-called warp distance that represents the effort needed to match one time sequence to the other. The bigger the effort (i.e., the warp distance), the more different the sequences are [149]. For our analyses, we used the DTW implementation for R by Giorgino [65] and computed the warp distances on a grid engine.

For the analyses, we used different combinations of touch screen data to create the time sequences. The combinations were based on the results by De Luca et al. who identified event time and xy-coordinates as essential parameters to distinguish between users [40]. We further included the distance between two fingers during drawing as an additional parameter. Table 5.2 overviews the combinations that we tested in our analyses and explains the parameters.

	T1	T2	T3	T4	T5	T6	T7	T8	T9
									
TP	171	131	183	179	166	161	161	174	175
TN	2158	2831	1999	2053	2634	2662	2028	2600	2535
FP	2042	1369	2201	2147	1566	1538	2172	1600	1665
FN	39	79	27	31	44	49	49	36	35
Sensitivity	81.4%	62.4%	87.1%	85.2%	79.0%	76.7%	76.7%	82.9%	83.3%
Specificity	51.4%	67.4%	47.6%	48.9%	62.7%	63.4%	48.3%	61.9%	60.4%
Accuracy	52.8%	67.1%	49.5%	50.6%	63.5%	64.1%	49.6%	62.9%	61.5%

Table 5.3: Overview of the best accuracy values for each sketch template.

Authentication Template Authentication templates are important as they decide whether an authentication attempt fails (i.e., the authentication attempt is not similar to the authentication template) or succeeds (i.e., the authentication attempt is similar to the authentication template). In order to create an authentication template for each sketch template, the data from the first session was used. Remember: each sketch template consisted of ten data sets (since each of them was drawn ten times during enrollment). The creation of the authentication template for each parameter combination, each user and each sketch template worked as follows: For each data set the warp distance to the other nine data sets was calculated and then averaged. The data set with the smallest average distance was then chosen as authentication template (i.e., it is the data set that is the closest to all other data sets).

In order to define a threshold T (i.e., warp distance) based on which an authentication attempt is considered as successful, the data sets were again compared to the authentication template, resulting in nine different warp distances. These distances were then used to calculate T . We tested different formulas (i.e., *mean*, *minimum*, *maximum*, *halfwayMeanMax*, *Standard deviation*) and found that, based on the resulting accuracy values, the *halfwayMeanMax* plus standard deviation worked best: $T = (\frac{\text{Mean}+\text{Maximum}}{2}) + \text{Standard Deviation}$. This observation is in line with previous work [40] and thus the *halfwayMeanMax* was used as the basis of our result reports in the next session.

Authentication and Attack The data from the second session was used to simulate fallback authentication after enrollment. For each parameter combination, each user and each sketch template, we compared the authentication attempts to the corresponding authentication template to decide whether the attempt was successful or not. In addition to this, we tried to attack each participant by using the authentication attempts by all other participants and compared them to the authentication template of the participant to be attacked.

Results

Altogether, we collected 1890 ($= 9 \times 10 \times 21$) data sets during the first session to create the corresponding authentication templates. The same amount of data was collected during the second session, but were used as authentication attempts and attacks, respectively.

Accuracy For each sketch template and each parameter combination the number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN) and accuracy values were calculated. Table 5.3 reports the results of the parameter combination that reached the best accuracy value for the corresponding sketch template.

Overall, the best accuracy was yielded for sketch template T2 when the factors *time*, x_2 , y_2 and *distance* were taken into account. This combination yielded an accuracy of 67.1% (1369 FP; 79 FN), with a success rate of 32.6% for attacks and a failure rate of 37.6% for authentication attempts.

In turn, sketch template T7 showed the worst results. Its highest accuracy value was only 49.6% (2172 FP; 49 FN). While the success rate for attacks was 51.7%, the failure rate was 23% for authentication attempts. Interestingly, this was the only sketch template that did not use the parameter *distance* to reach its highest accuracy. Instead, it took into account the parameters *time*, x_1 , x_2 , y_1 and y_2 .

Ease of Drawing After the enrollment of each sketch template, participants were asked to state how easy it was to draw the corresponding sketch with two fingers simultaneously, using a 5-point Likert scale (1=very difficult; 5=very easy). The opinions for most sketches were ambiguous, but for three of them, we observed a clear tendency. For example, sketch templates T3 and T6 were found to be difficult to draw by the majority of users as they had to move their fingers in opposite directions. For one of the two templates they even had to rotate their hand. In turn, sketch template T7 was easy to draw for the majority of participants. It consisted of two vertical and two horizontal strokes that could be drawn within two parallel strokes from top to bottom and from left to right.

Favorite Sketch The results of the ratings for ease of drawing were also mirrored in the sketches that users liked the most and the least, respectively. Sketch template T7 was the template that the majority of participants liked the most, while sketch template T6 was liked the least.

User Stories and Comments After the study, some participants reported that they were dissatisfied with the results of their sketches. They told us that most of the time it was not possible to draw a seamless sketch that did not have a gap when both of their fingertips met. Thus, the sketch differed from the provided template, resulting in frustration. To overcome this problem, some participants used only one finger to draw a line that closed the gap. Others gave up and drew the complete sketch with only one of their fingers.

5.1.6 Discussion

The study results clearly show that the proposed approach does not work for fallback authentication due to its bad usability and security properties (e.g., user frustration and the high number of FP/FN). But beyond that, we found some inspiring insights that should be taken into account when sketch-based authentication schemes are designed that use predefined sketch templates (but not two fingers simultaneously to draw them).

Usability Matters

Even if we had reached desirable security properties, the presented approach would have failed due to usability reasons that go beyond memorability issues. The qualitative ratings by participants were a prime example that user satisfaction is also an essential factor to be considered for fallback authentication (despite its infrequent use). Most participants found that sketching with two fingers simultaneously was frustrating and developed undesired behaviors (e.g., drawing the sketch with one finger only). This may have severe consequences: if users do not enroll as intended by the fallback scheme, the system may fail the user when it is actually needed (i.e., fallback authentication), for example, when they use one finger for enrollment, but two fingers for fallback authentication.

Distance between Two Fingers

The assumption that the distance between two fingers during sketching is distinct enough to distinguish between users was proven to be wrong. However, it is worth mentioning that the distance parameter was still a contributing factor in order to reach better accuracy values for most of the sketch templates.

Reasons why this was not sufficient to authenticate users and reject adversaries may be due to the limited complexity of the sketch templates or the form factor of smartphones. Since smartphone displays are small, the maximal possible distances between two fingers was not fully exploited by the participants. This could have been different on bigger devices (e.g., tablets).

Gap Closure

Participants in our study showed high perfectionism. They wanted their sketches to be as close to the template as possible. However, this was difficult to accomplish due to the use of two fingers.

When creating predefined templates for sketch-based authentication, it is important to provide users with a close-to-real experience, meaning that the template should reflect what is actually possible. With respect to the study templates, this could have been achieved, by including gaps to the sketch templates where both fingers were supposed to meet or by leaving the templates as is, but automatically connecting the gap during drawing instead.

Drawing Preferences

Most participants drew their sketches from top to bottom and from left to right. Overall, users preferred sketch templates that were simple and easy to learn. This is problematic as simple templates often do not convey enough data to distinguish between users. However, we assume that this problem is specific for 2-finger sketches and more complex sketches would not frustrate users when only one finger is used. But in case the problem is prevalent, one could think of including guiding instructions (e.g., arrows to indicate the direction of the stroke). In retrospect, ease of drawing (and thus drawing preferences) could have been positively improved, if our brainstorming had taken into account ergonomic characteristics of the human hand for the sketch templates.

5.2 Location-based Questions

For the design of fallback schemes, procedural skills were not the optimal solution. Therefore, we explore an alternative approach to traditional text-based security questions (see Section 2.3) that takes advantage of episodic memories. In the remainder of this section, we present the described approach in more detail: This encompasses the design of location-based questions, their implementation as well as evaluation. We further report the results of the evaluation to discuss their implications for the design of fallback schemes.

5.2.1 Approach

Our approach is based on a combination of location-based questions with map-based input. Although these questions are similar to traditional security questions where content is concerned (i.e., they are based on personal facts), they are inherently different as the answers are provided in the form of locations on a map instead of text. For example, the question “*Where was your first travel by plane?*” can be used for both location-based questions and traditional security questions. However, while the text-based answer for the latter can lead to lexical ambiguities (e.g., “*London, UK*” vs. “*London, United Kingdom*”), we assume that the use of map locations can overcome these problems as landmarks on the map (e.g., buildings or street crossings) can serve as memory cues.

Selecting a location on a map as password has already been proposed by previous work (e.g., [164, 168, 174]) and the answer input in our approach is inspired by one of them (i.e., GeoPass [174]). However, instead of memorizing arbitrary locations as primary passwords, we combine map-based input with location-based questions for cued-based recall, meaning that the questions are used to evoke episodic memories that, in turn, are associated with a particular location. This combination has not been explored before and has the advantage that no free recall is required. In particular in the context of fallback authentication, this is important to maintain memorability even when a long time between enrollment and authentication has passed.

The basic idea of our approach is as follows. During enrollment, users select or define a set of three questions for which they provide the answers in form of a map location. In case of password loss, users are, again, presented with the questions that they provided earlier. For each question, users are given three attempts to answer it. The answer is considered as correct when its distance to the original location (i.e., the location that users provided during enrollment) lies within a predefined threshold. In our approach, we use a threshold of 30 meters, as this was shown to be useful by Thorpe et al. [174]. Whether users are authenticated successful, depends on the number of questions that they answer correctly.

5.2.2 Question Design

A focus group with five participants was conducted to collect and discuss potential topics for the design of location-based questions. Participants were recruited over bulletin boards, mailing lists and personal communication. All of them were male and between 18 and 26 years old (average: 22 years). They were all students in natural sciences (e.g., computer science, physics and medical engineering).

Participants were accepted on a first come first serve basis, which could have influenced our question design. Since there is evidence from research that women have more vivid and precise autobiographical memories than men [87], our location-based questions are likely to be a lower-bound only.

For the focus group, participants were invited to our lab. The focus group started with a brief introduction to fallback authentication, the use of security question and an overview of our approach. After this, participants were encouraged to discuss the advantages and disadvantages of the presented approach. They were also asked to come up with topic ideas for the design of location-based questions.

Altogether, we collected a variety of promising topics, such as *childhood memories* (they lie far in the past and only few know about them), *travel memories* (they include all possible world locations and thus increase the answer space) or *first time memories* (they are special and memorable). Participants also mentioned *big events* (e.g., concerts) or *third persons* (e.g., childhood friends).

During the discussion, participants also addressed different question types. For example, since memories are very individual, participants raised concerns about the use of fixed questions due to potential applicability issues. The flexibility of open questions was also criticized as they have the risk that users define insecure questions. Overall, participants found guided questions to have the best trade-off between the two extremes: They set the frame for the question, but give users space for customization.

For the study, we tested all three question types to compare them to each other. Based on the insights from the focus group, we defined 22 fixed questions and 10 guidelines for guided questions (Table 5.4).

Fixed Questions

Whereto was your first travel by plane?	(5)
Where have you been camping for the first time?	(1)
Whereto was your longest travel so far?	(5)
Where was your first car accident?	(1)
Where is your favorite beach?	(3)
Where did you park for your driving test?	(1)
Where did your best friend from elementary school live?	(2)
Where did you injure yourself badly for the first time (e.g., broken leg)	(1)
Where was your first time at the sea?	(2)
Where did your best kindergarten friend live?	(0)
Where did you meet your best friend?	(2)
Where did you spend your first vacation?	(0)
Where did your first kiss take place?	(2)
Whereto did you drive in your first driving lesson?	(0)
Where have you been in a dangerous situation?	(2)
Where was your first party?	(0)
Where does a distant relative of yours live?	(1)
Where was your first breakup?	(0)
Whereto did you travel for your first school trip?	(1)
Where was your most embarrassing moment?	(0)
Where was your first job interview?	(1)
Where was your saddest moment?	(0)

Guided Questions

Please define a location-based question that ...	
... refers to a travel destination/vacation destination.	(7)
... refers to a personally experienced sport event.	(5)
... refers to an event in your childhood.	(4)
... refers to an event during your time at university/apprenticeship.	(4)
... refers to one of your party experiences.	(3)
... refers to something that you did for the first time.	(3)
... refers to an event during your time in school.	(2)
... involves another person.	(1)
... refers to one of your favorite places.	(1)
... refers to an experience that had a strong impact on your life.	(0)

Table 5.4: List of the 22 fixed questions and 10 guidelines used in the study. The values in brackets depict the number of times a question or guideline was selected by study participants. The original language was German.

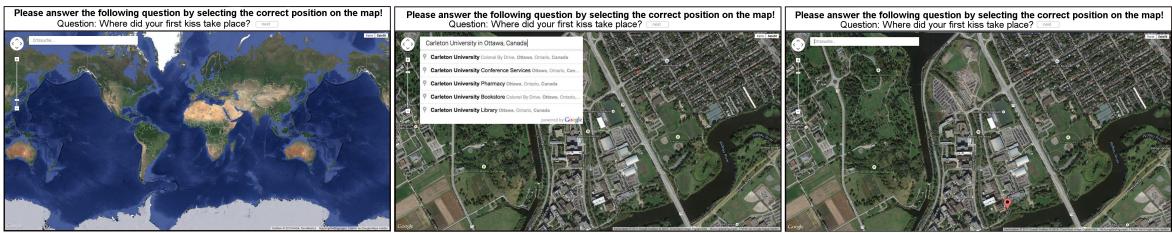


Figure 5.2: Screenshot of the user interface for the example question “*Where did your first kiss take place?*”. Answer provision always started with the world map (left). Users had the option to zoom in using either the mouse or to enter the required location in a search field (center). However, the answer had to be selected using the mouse to avoid facilitated guessing for adversaries (right).

5.2.3 Prototype

The study prototype was a simple web application that used the Google Maps API to retrieve location-based information (e.g., longitude and latitude). It further logged relevant information from user interactions (e.g., timestamps and selected/defined questions). The prototype had three main modes: enrollment, authentication and attack.

Enrollment

In enrollment mode users were asked to select/define three location-based questions. The overall procedure varied for the different question types: For fixed questions users were asked to select three questions from a list of 22 questions. For guided questions they were asked to select three guidelines from a list of ten guidelines. In addition to this, they were provided with three text fields so that they could define the corresponding questions. With respect to open questions, they were also provided with three text fields for question definition. After this was completed, users were consecutively shown the questions they had just selected/defined to provide the corresponding answers by selecting a location on the given map. Hereby, a location can be any geographical coordinates that, in turn, consist of latitude and longitude values. For each question, users could reposition their marker before submitting it with the corresponding save button.

Authentication and Attack

In authentication mode, fallback authentication was simulated. Users were consecutively shown the questions that they selected/defined during enrollment. For each question they were given three attempts to provide the correct answer on the map. Remember: An answer was considered as correct when its distance to the original location was within a threshold of 30 meters.

The attack mode was similar to the authentication mode. The only difference was that the answers were provided by potential adversaries and not the legitimate users.

Map Initialization, Zoom Level and Location Selection

In all three modes the map was always initialized at zoom level 2 and centered at the positions 0.0 / 0.0 (latitude/longitude) to display the complete world map. This was done to prevent biasing users during answer selection and to avoid hinting possible solutions to adversaries.

In order to zoom in, users could either use the mouse or the search field that we provided on the user interface to make it easier to find a specific region on the map. However, the answer selection had to be done using the mouse to make the answers more individual and harder to guess. Furthermore, the answer could only be submitted when the zoom level was higher than 16. Previous work by Thorpe et al. has shown the usefulness of this value [174]. In case the zoom level was too small, users were informed by a pop-up notification. Figure 5.2 shows an example work flow when providing the answer to a question.

5.2.4 Threat Model

We consider three types of threats for security evaluation: threats by close adversaries, threats by close adversaries that research possible answers on the Internet and strangers that also research on the Internet to conduct educated guessing attacks (see Section 4.3.2).

Close adversaries often have advanced knowledge about the user and are more likely to know the answers to autobiographical questions than other adversary types. The threat is aggravated when they use this knowledge to research possible answers on the Internet (e.g., on social networks).

With respect to strangers, their chances to guess the correct answers is $(\frac{1}{n})^x$, with x being the number of questions and n the number of selectable locations on a world map. However, in reality, n is assumed to be smaller when more targeted attacks are taken into account. For example, using Internet for research, strangers may narrow down the number of possible answers to a certain geographic location (e.g., the victim's home town).

Another threat to be mentioned are brute force attacks for which more sophisticated adversaries use automated processes to successively guess one possible answer after another. In order to encounter this threat, the number of guessing attempts must be limited.

5.2.5 User Study

The user study consisted of a short-term evaluation (three sessions within four weeks) and a long-term evaluation after six months. More details on the study and the different sessions are provided next.

Study Design

A between-groups design was used for the study to prevent biasing participants during the creation of questions, for example, to avoid that they define open or guided questions that

are similar to the ones that they encounter as fixed questions. The independent variable was *question type* with three levels (i.e., fixed, guided and open). The number of correct answers was used as dependent measure.

Study Procedure

Session I Participants were recruited for the study over bulletin boards, social media and personal communication. For the first session, participants were invited to our lab. They were asked to bring another person they were close to who acted as adversary (to attack their questions). Participants were given examples for close adversaries during recruitment (e.g., significant other, best friend). In order to avoid confusion, we will use the terms *users* and *adversaries* to distinguish between the two participant types.

To begin the study, users and adversaries were given a brief introduction to the topic of fallback authentication, the general idea of our approach and an overview of the study procedure. After this, adversaries were asked to leave the room, while users completed the actual task for the study group they were assigned to (i.e., fixed, guided or open). The task started with the selection/definition of three location-based questions (i.e., enrollment) and was followed by a short break during which users were shown a short distraction video with an approximate length of six minutes. This was done to conduct a memorability test shortly after enrollment. Users had three attempts per question to provide the correct answer. They were informed about the correctness of their answers. In case they were not able to provide the correct answer to a question within three attempts, they had to skip to the next question.

Once users completed their tasks, they were asked to leave the room. In turn, adversaries were invited back in and the same procedure was repeated, but without enrollment. In case adversaries were not able to guess the correct answer to a question within three attempts, they were given another three attempts to do some Internet research to find the answer.

Note that we allowed participants to act in both roles during the study in case they wanted to, meaning that they could take part as users as well as each other's close adversaries. We paid particular attention that these participants were assigned to different groups and that they, one after another, completed the user tasks (i.e., enrollment and memorability test) before the adversary tasks (i.e., attack). This was important to prevent biasing them during question selection/definition.

The first session was concluded by a questionnaire that had to be answered by users and adversaries. It contained questions about their demographics, the study task and their relationship to each other.

In addition to this, users were handed another form that asked them for personally identifiable information (i.e., first name, last name, date of birth and place of birth). This information is interesting for educated guessing attacks by strangers (see Section 4.3.2). We informed users about the purpose of this data collection and told them that the provision of this information was optional. However, none of the users refused to provide the needed information.

Session II + III The second and third session took place one week and four weeks after the first meeting, respectively. For both sessions, users were invited to our lab for another memorability test in which they had three attempts to answer the questions they had enrolled during the first session.

Long-Term Evaluation Another memorability test was conducted six months after the third session. This way, we were able to simulate a realistic fallback scenario in which a long time between enrollment and authentication had passed. In order to encourage users to participate and to spare them from long traveling times, the test was conducted over Skype.

Educated Guessing Attacks In addition to the attacks by close adversaries, we also considered educated guessing attacks. Two strangers were provided with a list of the users' location-based questions and personally identifiable information. They had two weeks to use this information for researching the answers to the questions on the Internet. For each question, they were instructed to select three locations they thought could be the answer. They were also asked to briefly state why they had chosen a specific location.

Incentives Users received 20 € gift vouchers as compensation for their participation when they completed the first three sessions (otherwise they received nothing at all). Another 5 € gift voucher was given when they took part in the long-term evaluation. Adversaries also received 5 € gift vouchers for their participation in the first session. This means that participants who took part in both roles received 25 € gift vouchers for the first three sessions.

Participants

General Demographics Altogether, 32 participants (15 female) took part in the experiment. While 28 of them acted as users as well as adversaries, there were two participants who acted as users only and another two who acted as adversaries only. Participants were between 17 and 55 years old (average: 26 years).

Most participants were students with different backgrounds (e.g., computer science, business, etc.). Others were in high school, employed (e.g., finance, administration, etc.) or retired.

Relationship When asked about their relationship to each other, most user-adversary pairs were best friends, good friends or partners/spouses. Overall there was good agreement between users and adversaries on the named relationships, but there were four cases in which the relationship did not match. For example, while one person treated the other as good/best friend, the other person considered them as acquainted friend/good friend only.

Participants were also asked to state how well they know each other on a 5-point Likert scale (1=not at all; 5=very well). Fourteen users-adversary pairs knew each other very well or well. One pair reported to know each other only a little. For another pair the ratings did not match. While one of them thought to know the other person well, the latter stated to know them a little only.

Strangers Two strangers (one female) who had no relation to the users in our study performed educated guessing attacks. They were 29 years and 33 years old. Both of them worked as research assistants at our university with a security background.

Results (Short-Term Evaluation)

Question Categories Altogether, 90 location-based questions were selected/defined in the user study (i.e., 30 fixed questions, 30 guided questions and 30 open questions).

Most users selected travel-related topics for one of their fixed questions (e.g., first flight or longest flight). This was followed by questions that involved another person (e.g., best friend or first kiss). Table 5.4 (top) shows how often each fixed question was selected.

Travel was also a popular topic for the guided questions. This was followed by guidelines about sport activities. Example questions that users defined included athletic first times (e.g., “*Where did I receive my first sports award?*” or “*Where did I run my first marathon?*”) and special travel memories (e.g., “*Where was your graduation trip?*” or “*Where did I spend my most beautiful summer when I was a child?*”). Interestingly, there was one question that referred to the future (e.g., “*Where is the place I want to be at least once?*”), while all other questions referred to the past.

For open questions, users covered different topics that, for example, included another person or animal (e.g., “*Where was my tomcat born?*”). Other questions were about special events (e.g., “*Where did I celebrate the victory against Argentina in 2010?*”), first times (e.g., “*Where did your first kiss take place?*”), travel (e.g., “*In which country did I get homesick?*”), education (e.g., “*Where was my final exam?*”) or personal preferences (e.g., “*Where can I eat my favorite food?*”).

Amount of Information in a Question During the review of the guided and open questions, we made an interesting observation: Some questions presume knowledge of more information than others. For example, in order to provide the answer to the question “*Where is the center of the route to my best childhood friend?*” five pieces of information are required: Who is the childhood friend? Where does the friend live? Where does the user live? Which route did the user take? Where is the center of this route?

Overall, 77% of the guided and open questions required only one piece of information (i.e., the location), 17% required two pieces of information (e.g., “*Where did I meet my best friend?*”) and 3% required three pieces of information (e.g., “*Where is my sister’s husband from?*”). Another 3% required four or more pieces of information.

Answer Distribution Figure 5.3 uses a world map to visualize the answers that users submitted during enrollment for their questions. Interestingly, most answers were clustered in one geographical area: Europe. In particular, most answers were located in Munich, Germany and its surrounding areas (i.e., the city in which the study took place).



Figure 5.3: Distribution of answers that users provided for their questions during enrollment.

Number of Correct Answers (Users) Users submitted answers to their security questions in three sessions and during the long-term evaluation (the results of the latter, however, will be reported in another section).

A two-way ANOVA that examined the effects of *question type* (i.e., fixed, guided and open) and *user type* (i.e., user, close adversary and stranger) showed that users were significantly better than adversaries in answering their questions for all three sessions (each with $p < 0.01$). No significant main effects were found for the factor *question type*. There were also no interaction effects. Thus, the results in the remainder of this section will report the total number of correct answers for all question types (and not for each of them separately).

An answer was considered as correct when the selected location was within a distance of 30 meters to the original location. In the first session, users answered 80 out of 90 questions (89%) correctly (Table 5.5). Most of them answered all their questions correctly and needed only one attempt most of the time. Although there were users that failed for some questions, none of them failed in all three questions. The main reason for failing a question was precision, meaning that users were close to the original location but outside the required threshold. Only in two cases were users far away from the the original location (> 1000 meters). The reason was that they had forgotten their answers. One of them remarked that she had selected a question/location that she did not associate with a strong memory.

After one week, users still answered 79 out of 90 questions (88%) correctly (Table 5.5). Similarly to the previous session, most of them answered all their questions correctly and most of them needed only one attempt. Interestingly, most users who failed some of their questions were the same users that already failed for these questions in the previous session. There was no user that failed in all three questions.

	Session I			Session II			Session III		
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
3 Attempts	1	1	1	0	0	0	1	0	1
2 Attempts	0	1	1	3	3	2	1	1	1
1 Attempt	26	25	25	25	24	25	23	24	21
3 Correct Answers	21		20			19			
2 Correct Answers	8		9			8			
1 Correct Answer	1		1			3			
0 Correct Answers	0		0			0			

Table 5.5: Overview of the number of attempts that users needed for the three questions (Q1-Q3) for each session as well as the number of correct answers by users per session.

In the third session (i.e., four weeks after the first one), answers to 76 out of 90 questions (84%) were recalled (Table 5.5). Most users answered all their questions correctly and were able to provide the correct answers within one attempt. Most users that had difficulties with their questions in the previous sessions still encountered these problems. However, two of them were able to provide the correct answers.

Number of Correct Answers (Close Adversaries) Adversaries were only required to submit their answers in the first session (as no memorability test was needed for them). They answered 6 out of 90 questions (7%) correctly. While some adversaries guessed the correct answers within one attempt, others needed two or three. Within the set of three questions, none of the adversaries had more than one correct answer.

Adversaries who failed to provide the correct answer after three attempts were given another three attempts to research it on the Internet. However, only two adversaries succeeded. Each of them answered one question correctly after research. Questions that adversaries were able to answer correctly were, for example, the questions “*In which street did my grandma live?*” or “*In which building was my first lecture?*”. The former question was guessed by a spouse, while the latter was guessed by a university friend.

Number of Correct Answers (Strangers) Most of the time, the two strangers failed in their educated guesses as, according to their reports, it was very difficult to research the answers despite the use of social networks and search engines. One of them was able to research the answer to 1 out of 90 questions (1%). The other had 2 out of 90 questions (2%) correct. Both of them needed between one and two attempts.

The questions that they answered correctly were “*Where was the first time I partied when I was a student?*”, “*Where did I meet my best friend?*” and “*In which building was my first lecture?*”. The correct answers were given due to different assumptions by the two strangers: For the last question, one stranger assumed the user to be a student at the university he was working at, and thus selected common buildings for classes. For another question, one of the

strangers assumed that the user had met the best friend in high school and selected different high school buildings in the home town of the user.

Answer Distances (Users) In case users answered a question incorrectly, we calculated the shortest distance of all three attempts to the actual location.

In the first session, users had 10 incorrect answers (11%). Seven of these answers were located 40-100 meters from the original location. Two answers had a distance of 300-600 meters and one answer had a distance of several kilometers.

After one week, users provided 11 incorrect answers (12%). Four answers were located 40-100 meters from the original location. The distance was 100-400 meters for five answers and over one kilometer for another two answers.

In the last session, the number of incorrect answers was 14 (16%). Nine answers had a distance 40-100 meters from the original location, two answers had a distance of 200-600 meters and three answers had a distance of several kilometers.

Answer Distances (Close Adversaries + Strangers) The same calculations were done for close adversaries and strangers.

Close adversaries gave 84 incorrect answers (93%), most of which were located multiple kilometers away from the original location (average: 440.4 kilometers). Twelve of their answers had a distance of 200-900 meters.

Similar observations were made for strangers. They answered 177 out of 180 questions (i.e., each of them attacked 90 questions) incorrectly (98%). For the majority of questions, their answers were multiple kilometers away from the original location (average: 1176.7 kilometers). In nine cases the distance was 300-800 meters and in another case the distance was 50 meters.

Authentication Time Time measurement began when users opened the first HTML page for enrollment/authentication using the corresponding start button. It ended when they submitted the last answer to their last question.

Users needed on average four minutes (min=75s; max=420s) for enrollment. Authentication took them on average 36s (min=12s; max=214s) in the first session, 45s (min=13s; max=225s) in the second session and 47s (min=13s; max=232s) in the last session.

Perceived Memorability In the first session, users were asked to state whether they think that they will be able to recall the answers to their questions after longer periods of time. They affirmed this for 78 out of 90 questions (87%), but negated it for three questions (3%). A neutral opinion was expressed for the remaining questions.

In the second session, users were asked how well they could recall the actual answers. For 73 out of 90 questions (81%) they did not have any difficulties at all. They needed some

time to think about the answers for 6 questions (7%) and they had forgotten the answers to 11 questions (12%).

The same question (as asked in the second session) was asked for the third session. No problems were reported for 71 out of 90 questions (79%). Some time to think about the answers was needed for 8 questions (9%) and the answers were forgotten for 11 questions (12%).

Guessability Users were asked to state how easy they think it is for different types of adversaries (i.e., close adversaries and strangers) to guess or research the answers to their questions. The ratings were given on a 5-point Likert scale (1=very difficult; 5=very easy).

While users thought that the answers to 34 of their 90 questions (38%) are not guessable by close adversaries, they assumed this to be possible for 35 of their questions (39%). The remaining opinions were neutral. They also believed that 48 of the 90 questions (53%) are not researchable by close adversaries, but assumed this to be possible for 27 of their questions (30%). They were neutral about the remaining questions. Furthermore, users thought that the majority of the questions is not guessable (89 questions; 99%) or researchable (85 questions; 94%) by strangers.

In turn, close adversaries were asked for each question whether they knew or guessed the answers. This was done to compare their assessment to their actual performance: For the majority of questions, 46 out of 90 (51%), close adversaries stated to have guessed the answer to the security questions. Only in one case the corresponding security question was answered correctly. For 38 questions (42%) they had some vague idea about the answer, but only in one case the corresponding security question was answered correctly. For six questions they were sure to know the answer and thus provided the correct answers to the corresponding questions. One of the reasons was, for example, an adversary that was part of the actual memory (e.g., joint vacation).

We further analyzed whether users and adversaries disagreed with respect to their guessability assessments. A disagreement was counted when users assumed their questions to be guessable by close adversaries, while the corresponding adversaries stated not to know the answers (or the other way around). A disagreement was found for nine questions, but only in one case did the user assume the question to be unguessable, while the corresponding adversary thought to know the answer to the question.

Accuracy In order to assess the memorability and security of the proposed approach, we calculated the number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN) as well as the accuracy values.

For the first calculation we considered authentication attempts by users and attacks by close adversaries only. For all three sessions, the best accuracy values were yielded when users were required to answer at least two out of three questions correctly (with three attempts per question). Table 5.6 gives an overview of the best accuracy values for these parameters. In the first and second session the accuracy values were 98.3% (0 FP, 1 FN), which, however,

	Session I	Session II	Session III	Long-Term
TP	29	29	27	20
TN	30	30	30	24
FP	0	0	0	0
FN	1	1	3	4
Accuracy	98.3%	98.3%	95%	91.7%

Table 5.6: Overview of the accuracy values for the first, second and third session when requiring users to answer at least two out of three questions correctly (with three attempts per question).

decreased to 95% (0 FP, 3 FN) for the third session due to an increase in the number of FN. However, no close adversary was able to authenticate. The same results were obtained when taking into account attacks by strangers only. The results are promising: None of the close adversaries or strangers were able to authenticate successfully, not even after trying to research the answers to the questions.

Results (Long-Term Evaluation)

Six months after the third session, another memorability test was conducted. Altogether, 24 out of 30 users from the preceding sessions took part in this long-term evaluation. The other six users did not respond to our invitation. To better understand the results that are reported next, it is important to note that all six missing users were able to authenticate successfully in the third session.

Number of Correct Answers After six months, users recalled 55 out of 72 questions (76%). While most of them needed one attempt to provide the correct answers, others needed two or three. There were eleven users who answered all their questions correctly. Nine users had two correct answers and four users had at least one question correct.

Overall, 17 out of 72 questions (24%) were answered incorrectly. Five answers were 40-100 meters from the original location. Another five answers had a distance of 100-800 meters and two answers had a distance of multiple kilometers.

Self-Assessment and Actual Performance When asked how well they recalled the answers to their questions after six months, most users stated that they did not encounter any problems at all (49 out of 72 questions; 69%). However, there were some users who needed some time to recall their answers (10 out of 72 questions; 14%) or who had forgotten the answers completely (13 questions; 18%).

The statements were in line with their actual performances. Only in two cases, users stated to have recalled the answer, but provided an incorrect answer instead. They were close to the original location, but not within the required threshold.

Accuracy Similarly to the preceding sessions, we used the accuracy metric for memorability and security analysis. When considering authentication attempts by users and attacks by adversaries only, the best accuracy value was 91.7% (0 FP, 4 FN) and required users to answer at least two out of three questions correctly (with two/three attempts for each question). The same parameters and accuracy values were identified when attacks by strangers were taken into account instead.

User Feedback Overall, users liked the presented approach. They found the use of location-based questions more secure and memorable than traditional security questions and would consider using them in a real-world deployment. However, users raised concerns about the fact that they were not aware of the required threshold of 30 meters. Most of them were confident that, if they had known about its existence, they would have been able to submit the correct answers. Another interesting observation was made by two users that reported about changes in the map sections for their security questions. Due to the different appearance of the map, they needed some time to orient themselves, as previous orientation points (e.g., buildings) were missing.

5.2.6 Discussion

The accuracy results from the short-term as well as long-term evaluation showed promising results in terms of security (i.e., none of the adversaries succeeded in their attacks). However, in terms of memorability, there were some users who failed fallback authentication, leaving room for the improvement of location-based questions, discussed next.

Question Type

Although we did not find any significant differences between fixed, guided and open questions, we made interesting observations with respect to the selection/definition for each question type.

For example, the use of open questions can lead to the definition of weaker (but not weak) questions that are similar to the ones found for traditional security questions (e.g., “*Where is my mother born?*”). Since this information can be researched from public records, potential adversaries may benefit from these kinds of questions (although it is still difficult to be close enough to the original location). The use of fixed questions can overcome these problems by excluding weak questions, but has the disadvantage that users cannot customize their questions to include a more personal notion (e.g., increasing the amount of information that is needed to answer a question) for better memorization and, in some cases, also for better security. All things considered, the use of guided questions seems to be the best trade-off between the different question types.

Topics of Questions

The user-defined questions in our study covered a variety of topics, such as travel, first times or special activities. Although these topics are similar to the ones found for traditional security questions, using them for location-based questions worked much better compared to how they were previously used (e.g., [72, 139]).

Nonetheless, the topics for the guided questions can be improved, for example, by encouraging users to define more complex questions that require the provision of multiple pieces of information (e.g., “*Please define a question that involves the center of two locations.*”). At the same time the amount of required information must be chosen carefully to avoid causing problems for users to come up with a question. In addition to this, it is also advisable to brainstorm about further topics for guided questions that take into account the thoughts of participants from different age groups to ensure the applicability of guidelines.

Memorability

Users were confident in their ability to recall the answers to their location-based questions when a long time between enrollment and fallback authentication has passed. This self-assessment was in line with their actual performances: Even after six months, the number of incorrect answers was few.

This shows that location-based questions work well in terms of memorability and have a promising potential for real-world deployments. We hypothesize that this is not only due to the great recall rates, but also due to the good self-assessment: Having a positive attitude towards an authentication system is a prerequisite to increase the motivation and time that users are willing to spend with the definition of questions for fallback authentication.

Answer Precision

Most of the time, when users gave an incorrect answer, the main reason was not related to memorability. Instead, errors were made due to the fact that users were not aware of the predefined threshold of 30 meters. They assumed the system to be more tolerant, resulting in imprecise selections that were close to the original location, but not within the desired threshold. Thus, it is important to inform users during enrollment about this threshold to improve their performance. Furthermore, most errors were made shortly after enrollment and then repeated in the following sessions. Therefore it is advisable to ask users to verify their answers during enrollment, for example, by re-entering the location. This way, users have the chance to adjust their position marker.

Guessability The majority of adversaries (i.e., close adversaries and strangers) had a hard time attacking the questions. Most of them had to guess the answers and even after research they were able to answer only few of the questions correctly. This shows that location-based questions work well in terms of security. Nonetheless, adversaries that share the same experiences that users had (e.g., joint vacation) are a serious threat. To prevent these adversaries from succeeding, it is important to define multiple questions from different categories.

Geographical Clusters Visualizing the locations that users provided during enrollment showed that most answers were centered around their home towns or universities. This is reasonable as they are likely to have spent a lot of time at these locations and associate many memories with them. However, this also means that potential adversaries can take advantage of this knowledge to perform targeted attacks by limiting the number of possible locations. This is exactly what the strangers from our study did. But even with this strategy, it was difficult for them to find the exact location.

5.3 Lessons Learned

In this chapter we have explored two types of autobiographical memories for the design of fallback schemes with static enrollments: drawing skills and episodic memories. While one memory type did not yield the expected results, the other had very desirable usability and security properties. The key insights are summarized in the following.

- **User Satisfaction.** Due to the infrequent occurrence of fallback authentication, we argued that some usability factors, such as user satisfaction, can, to some extent, be neglected for the design of fallback schemes (see Section 4.1.2). Although not a contradiction to this assumption, the results from both studies highlighted that user satisfaction matters after all and should not be forgotten completely. Fallback schemes that are disliked by users can lead to undesired behaviors, while schemes that meet their expectations are likely to be adopted (in particular when they require enrollment).
- **Behavioral Biometrics.** The use of implicit features from 2-finger sketches did not work to distinguish between users. Future implementations should focus on sketches that are drawn with one finger only. However, based on the experiences from our study and the insights from related work, we find that behavioral sketch-based authentication is a better fit for uses cases other than fallback authentication (e.g., continuous authentication) as they allow more flexibility in their design (e.g., implicit enrollments).
- **Location-Based Questions.** The use of location-based questions with location-based answer input is a great combination to overcome the problems of traditional security questions. Their usability and security properties are outstanding. We expect that, with some usability improvements, this combination would work great in a real-world setting.
- **Guidance.** One area of improvement is related to different notions of guidance. First, question design should focus on guided questions to encourage users to define complex questions that are harder to guess than other question types. Second, the overall system should provide users with more information about how it works (e.g., the use of answer thresholds) to improve the success rate of inputting correctly remembered answers.

6

Fallback Authentication Based on Dynamic Enrollments

Although the use of static enrollments has its advantages (see Chapter 5), it becomes problematic when a long time has passed between enrollment and fallback authentication. Information that once was up-to-date may have become outdated, making it difficult for users to recall the correct answers.

Dynamic approaches are a promising alternative to this as they do not require users to enroll in advance. Instead, information relevant for fallback authentication is collected implicitly, for example, through the user's interaction with the device.

This chapter explores the potential to exploit different categories of information found on smartphones for the design of three different prototypes for fallback authentication. This includes approaches using common smartphone activities (Section 6.1), icon arrangements on smartphones (Section 6.2) and installed apps (Section 6.3).

We found that app-related information categories are promising for the design of fallback schemes as they have a good balance between memorability and security. However, we also learned that the success of app-related information categories for fallback authentication depends on the design of the authentication task.

Overall, this chapter provides an overview of the advantages and disadvantages of different information categories for fallback authentication. It also highlights the factors that need to be taken into account when handling smartphone information for fallback authentication. In particular, this includes privacy aspects that have been rarely discussed in previous work about dynamic enrollments (Section 6.4).

Personal contribution statement: The content of this chapter is based on five bachelor theses by Stephan Thalhammer [173], Philipp Hauptmann [81], John-Louis Gao [62], Carina Saliger [150] and Manuel Demmler [45]. All theses were supervised by the author. Part of this work was published at the Conference on Human-Computer Interaction (CHI 2014 [74], CHI 2015 [75]) and the Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI 2015 [77]). Several co-authors were involved in these publications: Alexander De Luca, Manuel Demmler, Emanuel von Zezschwitz and Heinrich Hussmann. A detailed overview of task responsibilities can be found in the disclaimer.

6.1 Smartphone Activities

The use of dynamic enrollments relies heavily on data availability. Luckily, most smartphones are equipped with hardware and software sensors that collect all types of data (e.g., touch information or storage information). In order to exploit this data for fallback authentication, it is important to find data that is distinct enough to distinguish between users and that it is frequently updated over time to facilitate recall during fallback authentication.

Smartphone activities, such as calls or the use of apps, seem to be an interesting data source for the design of dynamic enrollments. Previous work has shown that individuals with different personality exhibit different smartphone usage patterns [30]. We hypothesize that these differences are distinct enough to be used for authentication. Das et al. [37], for example, used information about everyday smartphone activities for primary authentication. Their results were promising, but lacked a detailed security analysis that takes into account human adversaries. A similar idea was proposed by Dandapat et al., who used daily activity logs from phones, browsers or Facebook in the context of password sharing [36]. With a success rate of 95%, their results are promising. They also evaluated their approach with human adversaries, but did not take into account the closeness between the user-adversary pairs.

The research presented in this section is similar to the work by Das et al. and Dandapat et al. in the sense that we explore different types of smartphone activities [36, 37]. However, it is different as we use different types of human adversaries for evaluation. We further evaluate these activities in the context of fallback authentication, which imposes different requirements than primary authentication or password sharing (see Chapter 4).

The remainder of this section is structured as follows: Our research started with a brainstorming session to identify smartphone activities that are suitable for the design of security questions for fallback authentication (Section 6.1.1). This was followed by the implementation of a prototype for Android devices (Section 6.1.2) to evaluate these questions in terms of security and memorability (Section 6.1.4). The results were then used to improve the questions and to conduct a follow-up study to re-evaluate them (Section 6.1.5).

6.1.1 Question Design

The brainstorming session was joined by four smartphone users (one female). Participants were recruited over bulletin boards, social media and personal communication. They were between 23 and 25 years old (average: 24 years). All were students with a technical background. Each received 5€ gift vouchers as incentive for their participation.

During the session, participants were encouraged to name and discuss different smartphone activities that they thought to be suitable for the creation of security questions for fallback authentication. Altogether, participants came up with various ideas that can be grouped into seven categories: *SMS* (outgoing and incoming), *Call* (outgoing and incoming), *App*, *Music* and *Photos*. The categories are in line with the results from previous work (e.g., [36, 37]).

Category	Question + Time Span
SMS (out)	Whom did you text [yesterday last week last month]?
SMS (in)	Who texted you [yesterday last week last month]?
Call (out)	Whom did you call [yesterday last week last month]?
Call (in)	Who called you [yesterday last week last month]?
App	Which app did you use [yesterday last week last month]?
Music	Which artist did you listen to [yesterday last week last month]?
Photos	Which photo did you take [yesterday last week last month]?

Table 6.1: Overview of the 21 security questions of the first study.

For each of these categories, one question was created. In order to test how far we could go back in time, each question was combined with three different time spans: yesterday, last week and last month. Altogether, we ended up with 21($= 7 \times 3$) security questions for evaluation (Table 6.1).

6.1.2 Prototype

The study prototype consisted of two Android applications: one for logging and one for question generation.

Logging Application

Once installed, the logging application ran in the background of the users' device to avoid interrupting them during their everyday tasks. It logged all relevant information for the study and included information about incoming/outgoing calls (e.g., contact number, duration), incoming/outgoing text messages (e.g., contact number, length), app usage (e.g., name of app) and music (e.g., artist).

Since some of this data is highly sensitive, we ensured that our research complied with federal privacy laws. The logged data never left the users' device. We further informed users that they have the possibility to get a list of all information that was logged. We informed them about this in the installation instructions that we sent them with the study application. However, we never received a request to see this list.

Questions Application

The second application was responsible for creating the actual questions based on the logged information. The questions were generated in random order and each had four answer options, one of which was *none of them*. While one of the four options was the correct answer, the remaining ones were incorrect and were also based on the logged data (excluding the correct answer). All answer options were provided as text (Figure 6.1, left), except for photo-related questions for which the answers were represented by pictures (Figure 6.1, right).

We excluded popular apps, such as Facebook, for app-related questions to make it more difficult for adversaries to guess the correct answers. For this, we used the results of an online survey by Statista¹ that listed the 20 most used apps in 2012 (the year of the study).

6.1.3 Threat Model

We assume a person close to the user as potential adversary for the security analysis (see Section 4.3.2). The adversary has advanced knowledge about the victim so that providing the correct answers is not based on plain luck. The chance for a random adversary to guess the correct answers is $(\frac{1}{n})^x$, with n being the number of answer options and x the number of questions asked.

6.1.4 User Study

In this section, we provide an overview of the study design and study procedure, which is followed by the presentation of the results.

Study Design

A within-subject design was used for the study. The independent variables were *question category* (7 levels) and *time span* (3 levels), resulting in 21 dynamic security questions to be tested. As the dependent variable, we measured the number of correct answers.

The study consisted of two phases: a logging phase that lasted for four weeks and a one-hour lab study. For the latter, participants were required to bring another person to act as adversary. This person had to be someone they were close to. We gave participants examples of close persons (e.g., partners or family members) during recruitment, which happened over bulletin boards, social media and personal communication. In addition to this, it was a prerequisite to own an Android smartphone in order to participate in the study. In the remainder of this section we will refer to smartphone owners as *users*, while the term *adversaries* will be used for accompanying persons.

Study Procedure

For the first phase of the study, we contacted users by email to send them the study application with accompanying installation instructions as well as important information about the general procedure of the study. The logging phase lasted for four weeks, after which we reached out to users again to invite them and their corresponding adversary to our lab for the second phase.

¹ <http://de.statista.com/statistik/daten/studie/239434/umfrage/nutzeranteile-der-top-20-smartphone-apps-in-deutschland/> (*last accessed 29/04/2015*)

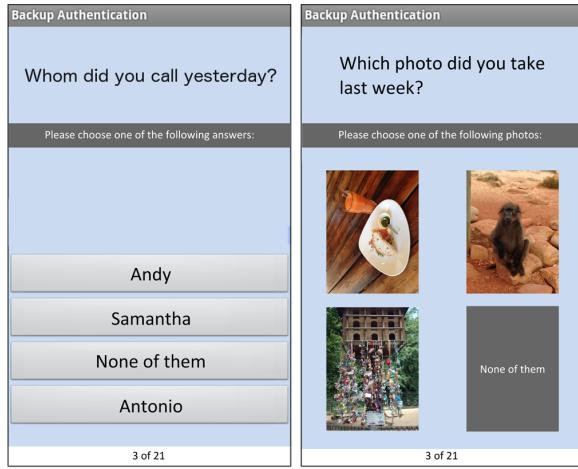


Figure 6.1: Example screenshots for two different questions. Each question has four answer options using either text (left) or pictures (right).

For each user-adversary pair, the lab study started with a brief introduction to the study procedure, after which the adversary was asked to leave the room. This was done to prevent the adversary from observing the user during the study task. In the next step, the study application was installed on the user's device and was followed by the actual task (i.e., answering the 21 questions). Then, the user was asked to fill out a questionnaire that covered demographic information, but also task-related items. Once the user was done, the adversary was invited back in and the same procedure was repeated.

Users received 20 € gift vouchers for their participation (10 € for the logging phase and 10 € for the lab study). In case they did not show up for the lab study, they did not receive any compensation. Adversaries also received 10 € gift vouchers for their participation.

Participants

General Demographics Altogether, 38 participants enrolled and took part in the user study: 19 users (7 female) and 19 adversaries (10 female). Users were between 20 and 31 years old (average: 26 years) and adversaries were also between 20 and 31 years old (average: 26 years).

Most users were students and research assistants with a technical background. When asked for their highest educational degree, fourteen of them reported a college degree and five of them reported a high school degree. Similar occupations were stated by adversaries. Twelve of them had a college degree as their highest degree, while seven had a high school degree as their highest degree.

Relationship Users brought different types of close persons to act as adversaries: eight brought their significant other, eight brought a friend and three brought an acquaintance. In three cases, the named relationships did not match. For example, while some users told us

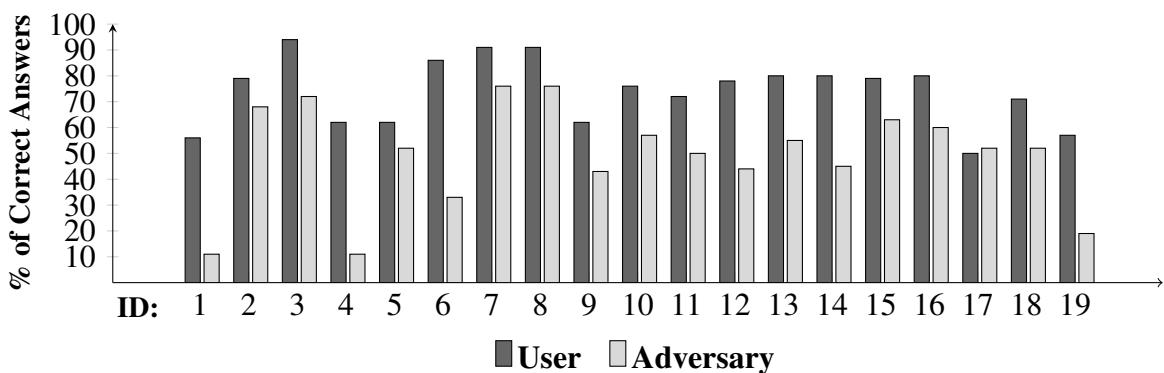


Figure 6.2: Overview of the number of correct answers per participant and their corresponding adversary (in %) for the first study.

to have brought a friend, the corresponding adversary stated to be an acquaintance only (or the other way around). However, the line between friends and acquaintances is often blurry. Overall, the named relationships matched well in both directions.

Results

Number of Correct Answers (Overall Performance) Altogether, we collected the answers to 744 questions (372 by users and 372 by adversaries), resulting in an average of 20 questions per user/adversary.

Users answered 274 out of 372 questions (73.7%) correctly. The best user had 94.4% of correct answers, while the worst user only 50%. The average success rate was 73.9%. In turn, adversaries answered 191 out of 372 questions (51.3%) correctly. The best adversary had 76% of questions correct. The worst adversary answered only 11% of them correctly. The average success rate was 51.3%.

Some users had a success rate of over 90% and were very good in answering their questions (e.g., users 3, 7 and 8). But at the same time, their adversaries performed well, too. They had over 70% of correct answers. One of the reasons for this was related to the relationship between users and adversaries. For example, users 7 and 8 participated as users as well as each other's adversaries in the study. They were best friends who communicated on a daily basis using their smartphones. For some questions (e.g., “*Whom did you call yesterday?*”) the adversary herself was the answer to the question. In turn, some users seem to know very little about themselves (e.g., users 1, 17 and 19). They had less than 60% of their questions correct. Figure 6.2 gives an overview of the individual performance of users and their corresponding adversaries.

Number of Correct Answers (per Category) Users were best at answering questions about outgoing text messages as well as incoming text messages (Table 6.2). They had 93% of correct answers for *SMS (out)* and 79% for *SMS (in)*. However, the corresponding

	Question Category						
	SMS (in)	SMS (out)	Call (in)	Call (out)	App	Music	Photos
User	79.0%	93.0%	77.2%	71.9%	71.4%	62.2%	56.9%
Adversary	61.4%	64.9%	47.4%	50.9%	35.7%	48.7%	47.1%
Difference	17.6%	28.1%	29.8%	21%	35.7%	13.5%	9.8%

Table 6.2: Overview of the number of correct answers (in %) by users and adversaries for each question category as well as the difference between them.

adversaries were good in answering questions from those categories as well. They had a success rate of 64.9% for *SMS (out)* and 61.4% for *SMS (in)*, respectively.

The performance of adversaries was worse for the other categories. They had less than 50% of correct answers for *Call (in)*, *App*, *Music* and *Photos*. However, for the latter two, users achieved only a success rate of 62.2% and 56.9%, respectively. The best trade-off was found for the question category *App*. While users were able to answer 71.4% of the questions correctly, adversaries achieved only 35.7%.

Number of Correct Answers (per Time Span) Users answered 87.3% of yesterday's questions correctly. This was more than for questions about last week (71%) or last month (62.3%). We did not observe any tendencies with respect to adversaries. They had around 50% of correct answers for the three time spans.

Self-Assessment and Actual Performance After the study task, users and adversaries were asked to estimate the number of correct answers they thought to have given. The majority of them tended to overestimate their performance. On average, the difference between the number of estimated answers and the number of correct answers was 3 (min=0; max=8). In turn, the average difference for adversaries was 4 (min=0; max=17).

Guessability Users were also asked to state how easy they think it is for an adversary to guess the correct answers to certain questions. While they perceived questions about *Music* and *App* as safe categories, there was no clear opinion for the remaining categories. We also did not observe any tendencies with respect to the different time spans that users found less or more secure.

Most users thought that the presented approach is safe against strangers. However, it was assumed that close persons or even acquaintances are able to answer at least some of the questions correctly.

User Stories and Comments The qualitative feedback that we received from our users at the end of the study was positive. Most of them described our approach as entertaining and fun. One user, for example, was particularly excited to learn about her own smartphone behavior. However, some participants raised concerns with respect to data privacy. One user-adversary pair talked to us at the end of the study and said: “We just broke up – just kidding,

but your study could destroy relationships. You are revealing information that you actually want to protect." They were referring to answer options that revealed communication details or photos. Two other participants had their birthday during the logging phase and thus had difficulties in answering questions related to communication categories (e.g., "I would have been better in answering questions about last month, if it wasn't my birthday. I received so many text messages").

6.1.5 Follow-Up Study

Based on the insights from the previous study, the questions design was improved for further evaluation. The goal was to test the security of the different question categories with respect to different types of adversaries (i.e., close adversaries and acquainted adversaries). In addition, we wanted to evaluate other app-related questions as they yielded promising results in the previous study. We further wanted to see if the privacy concerns for photo-related questions could be reduced when blurring them before showing.

In summary, the overall study design for the follow-up study remained similar as for the previous study, but tested slightly different set of question categories: The question category *Music* was removed due to the difficulties that users had had in answering them. Despite a similarly bad performance for photo-related questions, we kept them to test whether blurring photos before showing can reduce privacy concerns. Furthermore, we included another question category called *App Install*. The communication-based question categories were not modified and were kept to re-evaluate them with respect to different types of adversaries. Last but not least, we removed the time span last month as users were better in answering questions about yesterday or last week. An overview of all questions is given in Table 6.3. A brief recap of the study design, study results and their discussion is given next.

Study Design and Study Procedure

The follow-up study was similar to the first study, using a within-subject design with the independent variables *question category* (seven levels) and *time span* (two levels). This resulted in the evaluation of 14 security questions. Again, the study consisted of a logging phase and a lab study. However, this time users were required to bring two types of adversaries: a person that they are close to and a person that they are acquainted with. This allowed us to evaluate the security of the questions with respect to different types of adversaries.

The study procedure was identical to the first study, but took into account the acquainted adversary who had to wait outside the study room until the user and the close adversary had, one after another, completed the study task.

Participants

General Demographics Users and adversaries who took part in the previous study were not allowed to participate again. Altogether, we recruited 18 users for the follow-up study,

Category	Question + Time Span
SMS (out)	Whom did you text [yesterday last week]?
SMS (in)	Who texted you[yesterday last week]?
Call (out)	Whom did you call [yesterday last week]?
Call (in)	Who called you [yesterday last week]?
App	Which app did you use [yesterday last week]?
App Install	Which app did you install/update [yesterday last week]?
Photos	Which photo did you take [yesterday last week]?

Table 6.3: Overview of the 14 security questions of the follow-up study.

but ended up with 11 users (5 female) only. They were accompanied by 11 close adversaries (3 female) and 11 acquaintances (2 female). Reasons for the high dropout rate were, for example, the vacation season (e.g., users did not show up for the lab study) or technical issues (e.g., question application could not be installed; see Section 6.1.2).

Users were between 19 and 33 years old (average: 24 years). Close adversaries were between 19 and 33 years old (average: 23 years) and acquainted adversaries were between 19 and 58 years old (average: 27 years). When asked about the highest educational degree, two users reported a college degree, while nine users stated to have a high school. Similar backgrounds were reported by close adversaries. Six of them had a college degree as highest degree, while five had a high school degree. With respect to acquainted adversaries, five of them had a college degree as highest degree and six had a high school degree.

Relationship For the role as close adversary, three users brought their significant other, five brought a very close friend and one brought her brother. All named relationships between users and adversaries matched well. In addition to this, users and adversaries were asked to rate how well they know each other on a 5-point Likert scale (1=not at all; 5=very well). All users and adversaries stated to know each other well or very well.

Eight participants stated to have brought a college friend, high school friend or colleague to act as acquainted adversary. Two brought a friend, while another two brought a friend of a friend. There were no contradictions with respect to the named relationships. When asked to rate how well they know each other, most users and acquainted adversaries stated that they did not know each other well.

Results

Number of Correct Answers (Overall Performance) Altogether, we collected the answers to 462 security questions (154 by users, 154 by close adversaries and 154 by acquainted adversaries). Users answered 135 out of 154 questions correctly (87.9%). The best user achieved 92.9% and the worst user 78.6%. The average success rate was 87.7%.

Close adversaries answered 84 out of 154 questions correctly (54.6%). The best close adversary achieved 85.7%, while the worst had 21.4%. The average success rate was 54.6%.

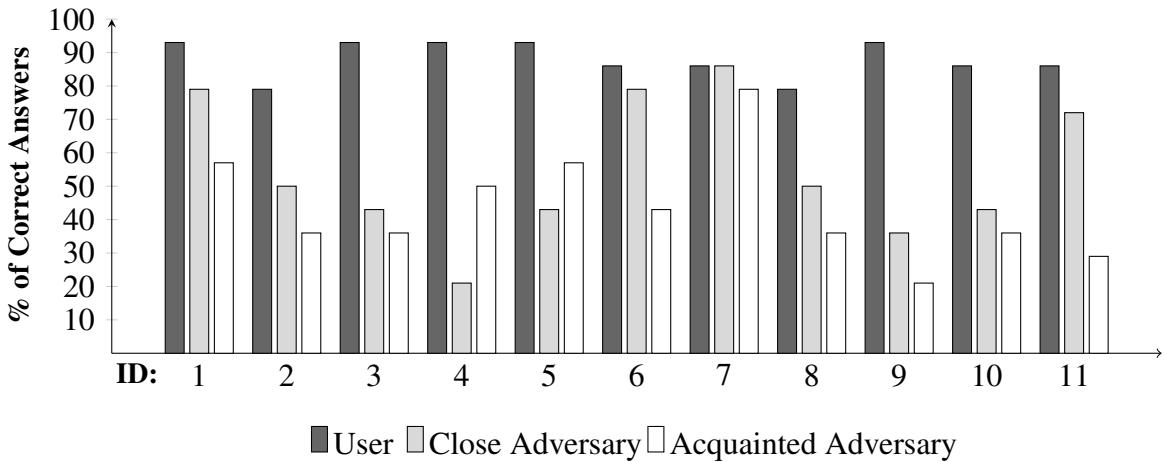


Figure 6.3: Overview of the number of correct answers per participant and their close and acquainted adversaries (in %) for the follow-up study.

In turn, acquainted adversaries provided correct answers to 67 out of 154 questions (43.5%), with 78.6% being the best and 21.4% the worst percentage of correct answers. The average success rate was 43.5%.

Figure 6.3 overviews the individual performance of each user and adversary (close and acquainted). Overall, users were good in answering their questions and made only few errors (1-3). Five users had over 90%, four users had over 85% and two users had over 75% of their questions correct. However, some users were easier to attack than others. For example, users 1, 6, 7 or 11 were good in answering their questions, but the performance of their close adversaries was good as well. In turn, users 3, 4, 9 or 10 were harder to attack. While they had a high number of correct answers, their close and acquainted adversaries reached only a low percentage of correct answers.

Number of Correct Answers (per Category) Users were good in answering most of the communication-related question categories (Table 6.4). For example, they answered all questions for *SMS (in)* correctly. However, close and acquainted adversaries were good in answering these questions as well. They reached 68.2% and 63.6%, respectively.

Adversaries performed worse on other question categories (i.e., *Call (out)*, *App*, *App Install* and *Photos*) and achieved success rates of around 50%. The best trade-off was found for the question category *App Install*. The difference between users and adversaries was 50% for close and 59% for acquainted adversaries. Although the differences were also promising for some other question categories (e.g., *Photos*), they suffered from problems, such as the revelation of private information. For example, users stated to feel uncomfortable having photos as answer options (even if they were blurred).

	Questions Category						
	SMS (in)	SMS (out)	Call (in)	Call (out)	App	Install	Photos
User	100 %	90.9%	81.8%	86.4%	72.7%	90.9%	90.9%
Close Adv.	68.2%	59.1%	68.2%	45.5%	50%	40.9%	50%
Acquainted Adv.	63.6%	36.4%	36.4%	36.4%	54.6%	31.8%	45.5%
Difference 1	31.8%	31.8%	13.6%	40.9%	22.7%	50%	40.9%
Difference 2	36.4%	54.6%	45.5%	50%	18.2%	59.1%	45.7%

Table 6.4: Overview of the number of correct answers (in %) by users and adversaries (close and acquainted) for each question category as well as the difference between them.

Number of Correct Answers (per Time Span) When answering the questions, users made very few errors. Interestingly, users made at most one error for questions about yesterday. This kind of error was always made for questions from the category *App*.

Altogether, users gave 94.8% of correct answers for questions about yesterday and 80.5% of correct answers for questions about last week. Close and acquainted adversaries were also better in answering yesterday's questions, but the differences to last week's questions were minimal. Close adversaries achieved 55.8% for questions about yesterday and 53.3% for questions about last week. Acquainted adversaries had 44.2% and 42.9%, respectively.

An ANOVA was conducted using the between-groups factor *user type* (i.e., user, close and acquainted adversary) and the within-factor *time span* (i.e., yesterday and last week). There was no significant main effect for the factor *time span*, but a significant main effect was found for *user type* ($F(2, 30) = 24.1; p < 0.001$). The post-hoc test showed that users were significantly better in answering their questions than close ($p < 0.01$; *Bonferroni corected*) and acquainted adversaries ($p < 0.01$; *Bonferroni corected*). No differences were found between the two adversary types.

Self-Assessment and Actual Performance Users were good in assessing their own performance. The difference between actual performance and estimation was on average one question (min=0; max=2). The average distance for adversaries was three questions (min=0; max=8).

Guessability Questions about communication-related activities or the category *Photos* were considered as insecure by the majority of users. Only questions about yesterday's *App* were found to be secure, while the opinions for the remaining question categories varied among users.

Accuracy So far, we have only reported the number of participants who answered individual questions correctly, but fallback systems usually consist of a combination of multiple questions to enhance security. To identify the best combination within $n = 14$ questions (as we had 14 questions in the study), we calculated the accuracy values for all possible combinations (i.e., $\sum_{k=1}^n \binom{n}{k} = 16383$).

One of the best combinations consisted of three questions (i.e., “*Which app did you install yesterday?*”, “*Whom did you call yesterday?*” and “*Whom did you text yesterday?*”) and required users to answer all of them correctly in order to authenticate successfully. With this, we achieved an accuracy of 95% (1 FP, 0 FN) when only attacks by close adversaries were considered. If only attacks by acquainted adversaries were taken into account, the accuracy was even 100% (0 FP; 0 FN).

However, with only three questions (and four answer options per question), the chances for a random adversary to authenticate successfully are $(\frac{1}{4})^3 = 0.016 = 1.6\%$ and thus much higher than the chances to guess a four-digit PIN ($\frac{1}{10000} = 0.0001 = 0.01\%$). Luckily, there were several other combinations that yielded the same accuracy values, but consisted of more questions (e.g., “*Which app did you install yesterday/last week?*”, “*Which app did you use last week?*”, “*Whom did you call yesterday?*”, “*Who called you last week?*” and “*Who texted you yesterday/last week?*”).

6.1.6 Discussion

The evaluation of different smartphone activities for the design of fallback schemes showed promising results in terms of usability and security. The removal of questions categories that were hard to answer (i.e., *Music*) and the inclusion of other question categories (i.e., *App Install*) seemed to have positively influenced the performance of users and their self-assessment. However, the results also identified the challenges when exploiting personal smartphone data. The key lessons from both studies are summarized in the following.

Limited Device Usage

Users performed very differently in our study. We had users who were good in answering their questions, but we also had users who knew very little about themselves. Interestingly, the latter type of user was also very hard to attack. We assume that these users spend only little time with their smartphones and, in addition to this, have atypical usage patterns that are difficult to guess (e.g., the use of special apps). The main problem with these users is that there might not be enough usage data to generate the security questions for fallback authentication. One could think of creating more security questions within a question category for which enough data is available, but in the last resort, an alternative for fallback authentication is needed. This is essential to prevent adversaries from succeeding in their attacks by always answering with *none of them* when the smartphones had been idle for some time.

Time Spans

Overall, users found it easier to answer questions about yesterday than last week. This was also mirrored in their actual performance and is reasonable as recent activities are more present in one’s memory. Nonetheless, we also found that some questions worked better when combined with longer time spans. This included, for example, questions from the

category *App*. A possible explanation is that apps are often used habitually (e.g., checking emails) and part of an automated routine for which it is easier to tell if it happened some time during the week than to answer if it occurred the day before.

Special Events

The occurrence of special events made it difficult for users to answer certain questions. For example, two participants received too many text messages for their birthday so that they could not answer the question “*Who texted you last month?*”. Therefore, unusual events need to be taken into account, for example, by outlier detection. In case an event happens more frequently than usual, the corresponding question should be skipped during fallback authentication. Another possibility could be the inversion of the question. Instead of asking “*Who texted you yesterday?*”, one could ask “*Who did not text you yesterday?*”. However, we have to keep in mind that the latter approach could cause social tension by pointing out that some expected event did not happen.

Question Categories

One of the key insights from both studies was that the best option in terms of usability (i.e., memorability) is not necessarily the best option in terms of security and the other way around. For example, although users were good in answering communication-related questions, adversaries did a good job as well. Instead, the best trade-off was found for app-related questions. Furthermore, most users perceived these types of questions as personal, but not too private so that they had less concerns when revealing some of their used apps to potential adversaries. A contrary example was the use of photos or contact names as answer options that most participants described as privacy intruding.

Similar to the category *App*, *Music* was also described as a safe question category. However, most users had problems in answering them. One of the reasons might be the way music on mobile devices is consumed. Listening to music is a passive activity that often happens on the go or while doing other activities. Thus, it is difficult to recall which music one had listened to at a specific point in time. In addition to this, many users stream their music from different services (e.g., Spotify) making it more difficult to log the data for question generation.

Security

The redesign of the questions from the first study increased the number of correct answers given by users during the follow-up study. It also influenced the performance estimation positively, meaning that users were confident in the answers they had given. This is important to offer users a close to real experience during fallback authentication, making the overall process more enjoyable and less frustrating. However, close and acquainted adversaries had a good self-assessment as well, indicating that they know well which questions they can or cannot answer. This is dangerous as adversaries (in particular close ones) could try to spy on their victim to collect the missing pieces.

It is also important to use multiple questions for fallback authentication to make it more difficult for them to guess or observe the correct answers. One of the best combinations that we found consisted of seven questions with a comparable security level to a 4-digit PIN. However, the combination is very strict, allowing users to make at most one error. Furthermore, the combination included communication-related questions that users found as privacy-intruding. Although this does not mean that these questions should not be used at all, it seems to be advisable to find more app-related questions to complement the two categories *App* and *App Install*.

6.2 Icon Arrangements

In this section, we explore the potential of icon arrangements as another app-related option for fallback authentication. Most smartphones have so-called home screens: Dedicated areas for customization that, for example, allow users to choose their background images or to arrange home screen items on a predefined grid. Home screen items include app icons, folders or widgets (i.e., apps that run on home screens).

Since different strategies exist for home screen organization (e.g., arrangement by frequency or color) [14], we hypothesize that icon arrangements should be distinct enough to be used for fallback authentication. While adversaries are unlikely to know the exact layout, we assume that users implicitly learn their icon arrangements through their daily smartphone interactions and are able to reconstruct their home screen layout for authentication.

This assumption is supported by earlier experiments that identified the storage of spatial information in long-term memory as a by-product of the interaction with objects [115]. In the context of desktop computers, similar results were found by Ehret et al. [54]. There is evidence that this could be also the case on smartphones. In a study by Gustafson et al., users were able to name the location of 68% of their apps by heart [71].

In the remainder of this section, we provide details about our approach: We summarize the results of a brainstorming that inspired the implementation of our study prototype. We further report the results from two consecutive user studies that were used to evaluate icon arrangements for fallback authentication and discuss the key findings from these studies.

6.2.1 Approach

A brainstorming with five participants (one female) was conducted to collect ideas for home screen-related authentication. Participants were recruited over bulletin boards, social networks, email or personal communication. They were between 20 and 26 years old (average: 22 years) and had a background in computer science. All of them owned a smartphone and were familiar with the concept of home screens. Each of them received 5 € gift vouchers as incentive for their participation.



Figure 6.4: Example screenshots for *Puzzle Tiles*, *Widget Space* and *App Selection*. The left screens are original home screens of users. The right screens are the corresponding solutions for each approach. For the first approach, users had to place different app icons onto the correct fields of the home screen. For the second approach, users were asked to select the fields that were occupied by Widgets. For the third approach, users were asked to select the app icons that are not correctly positioned on their home screens.

The brainstorming started with a brief introduction to fallback authentication and was followed by the collection of ideas that participants were encouraged to express openly. The results included a variety of ideas, such as security questions about the number of used home screens, the identification of one's home screens from a set of different layouts or the selection of one's background image from a set of images. Since the listed ideas had a very limited question space or potential privacy risks (i.e., users often have personal photos set as their background image), we concentrated on other ideas that participants named during the brainstorming. The corresponding ideas are described next.

Puzzle Tiles The idea for this approach is based on a puzzle metaphor. Each piece of the puzzle is a home screen item that needs to be positioned correctly on the home screen grid (Figure 6.4, left). Users are authenticated when they position a certain number of their home screen items correctly.

Widget Space This approach focuses on widgets only. In order to authenticate, users are required to mark all fields on a home screen that are occupied by widgets. For example, for a home screen with widgets at the top left corner (size: 4×2) and at the bottom right corner (size: 1×1), users have to mark all fields in the first two rows and the last field at the bottom right (Figure 6.4, center). If authentication succeeds or fails, depends on the number of correctly marked or unmarked fields.

App Selection This approach consists of a pre-assembled home screen with 16 icons. While some of these icons are part of the user's original home screen, others are not. In order to authenticate, users have to mark the app icons that are part of their original home

screen and that are located at the correct position (Figure 6.4, right). Authentication success depends on the number of correctly identified apps.

6.2.2 Prototype

All three approaches were implemented for smartphones running Android 4.2 (Jelly Bean) or higher. While all of them were similarly structured by using a 4×4 home screen grid (dock excluded), their implemented features varied.

Puzzle Tiles uses an app drawer that contains all home screen items that need to be positioned for authentication. Each item within the drawer can be dragged to the home screen and dropped on one of the 16 fields. In case placement is not possible (e.g., widget is too wide to be placed at the selected position), users receive visual feedback.

For *Widget Space*, users can mark or unmark each of the 16 fields on the grid by simply touching them. The state of each field is color coded: unmarked fields are grey, while marked fields are blue.

For *App Selection* the home screen fields are pre-filled with 16 different app icons. Some of these icons are part of the user's original home screen, while others are retrieved from an app library that consists of the 20 most downloaded and free apps from the Google Play Store (at the time of this work). The number of library apps to be included is a random value between 0 and 16. Similar to *Widget Space*, users can mark and unmark each of the app icons. Again, the state of each app icon is color coded: apps with a grey background are unmarked, while apps with a blue background are marked.

In order for the three prototypes to work, information about the home screen items needs to be extracted. Since this information is managed by the launcher (i.e., the main view of the device), we parsed the launcher's MySQL database to retrieve various information (e.g., path to app icons, app names, xy-coordinates or the dimension of widgets). It should be noted that not all of this information is mandatory. For example, in most cases there is no information about the name of a widget.

6.2.3 Threat Model

Similar to the previous studies, we assume an adversary that is close to the user and in possession of the user's device. The adversary tries to circumvent primary authentication by exploiting the fallback scheme (see Section 4.3.2).

The chances for a random adversary to guess the correct answers varies for each approach. For *Puzzle Tiles* they depend on three factors: a) the number of items to be placed, b) their position on the grid and c) their dimensions. In the worst case, the chances are 1 (i.e., when a widget with the size 4×4 is placed). In the best case, the items to be placed are all apps and folders. With n items and x attempts, the chances then are $\sum_{i=0}^{x-1} 1 / \left(\frac{16!}{(16-n)!} - 1 \right)$. Assuming

a home screen with $n = 2$ items and $x = 3$ attempts, there are $16!/(16 - 2)! = 240$ possible ways to place them on the grid and thus the chances are $\frac{1}{240}$ to guess the correct answer within the first attempt. With each incorrect guess (and additional attempt), the number of possible solutions is reduced: The chances to guess the correct answer is $\frac{1}{239}$ for the second and $\frac{1}{238}$ for the third attempt. This results in an overall probability of $\frac{1}{240} + \frac{1}{239} + \frac{1}{238} = 0.013$.

For *Widget Space* and *App Selection*, there are $a = \sum_{i=1}^{16} \binom{16}{i} = 65535$ possible solutions. Therefore, the chances for a random adversary to guess the correct answers is $\sum_{i=0}^{x-1} \frac{1}{a-i}$, with x depicting the number of attempts.

6.2.4 User Study

In this section we describe the study design and procedure to evaluate the three approaches. We further present the results of this study.

Study Design

A mixed study design was used for evaluation. The independent variables were *approach* (three levels) and *screen type* (two levels). The approaches were evaluated using a between-groups design to avoid learning effects, for example, to prevent users from learning their home screen organization from previous tests. In turn, a within-subject design was used for the different screen types. Furthermore, their order was counterbalanced to minimize learning effects. Altogether, we tested two screens: the screen that users see first when unlocking their device (i.e., main screen) and the screen with the most app icons, widgets and folders (i.e., secondary screen).

Study Procedure

Participants were recruited over bulletin boards, social media, email and personal communication. Participants who took part in the study were required to own a smartphone and to bring another person that knows them well to act as adversary. In order to avoid confusion, we will use the terms *user* for smartphone owners and *adversary* for accompanying persons.

The study was conducted in our lab and started with a brief introduction to the general procedure. After the introduction, adversaries were asked to leave the room, while users stayed in the room. Users were assigned to one of the three study groups (using a Latin Square). We further installed the study application and let users complete their study task. Users ended the study with a questionnaire about demographic information and task-related questions. After this, adversaries were invited back in and the same procedure was repeated.

The study lasted about 30 minutes. Users and adversaries each received 5€ gift vouchers for their participation.

Participants

Altogether, 36 participants took part in the experiment (6 users and 6 adversaries per group). However, for *App Selection* the data of one user-adversary pair had to be removed due to incomplete data from the launcher's database. All users owned an Android device with a launcher that uses a 4×4 home screen layout (dock excluded). This was a prerequisite to participate in the study.

General Demographics Six users (2 female) and six adversaries (all male) tested *Puzzle Tiles*. Users were between 20 and 29 years old (average: 23 years). Adversaries were between 20 and 26 years old (average: 23 years).

Widget Space was also tested by six users (all male) and six adversaries (2 female). Users were between 17 and 29 years old (average: 22 years), while adversaries were between 18 and 24 years old (average: 21 years).

Five users (2 female) and five adversaries (1 female) tested *App Selection*. Users were between 20 and 25 years old (average: 22 years). In turn, adversaries were between 20 and 34 years old (average: 26 years).

The majority of users and adversaries had a background in natural or engineering sciences (e.g., computer science or physics). Others came from areas like teaching or psychology. The distributions were similar over all three groups and also similar between users and adversaries.

Relationship Users and adversaries were in good agreement on their relationships. For *Puzzle Tiles*, five pairs were friends and one pair were colleagues. For *Widget Space*, four pairs were friends, one pair were colleagues and another pair were roommates. With respect to *App Selection*, two pairs were in a romantic relationship, another two were friends and one pair were colleagues.

Home Screen Organization All users stated to personalize their home screens using different strategies that, for example, are based on usage frequency, app similarity or aesthetic factors. These reports are in line with previous research (e.g., [14]). Users also reported that the re-organization of apps happens infrequently. Most of the times this is done when new apps are installed (e.g., about once a month or, in some cases, even once a week).

Device Sharing When asked whether they had used the victim's device before, 13 out of 17 adversaries affirmed this and reported that sharing was most of the time limited to a few minutes. This information is interesting to see whether adversaries had had the chance to learn about the user's icon arrangements before the study started.

Results

Number of Home Screen Items The number of home screen items varied from user to user. For the main screen, they had on average 7 items (min=3; max=13). For the secondary screen, the average number was also 7 (min=1; max=13).

Number of Correct Answers (Puzzle Tiles) A correct answer was counted when an app, widget or folder was correctly positioned on the grid. Table 6.5 (columns 1-2) reports the percentage of correct answers. Users gave 82.1% (min=28.6%; max=100%) of correct answers for the main screen and 65.2% (min=33.3%; max=100%) for the secondary screen. In turn, adversaries gave 43.6% (min=0; max=100) and 15.2% (min=30.7%; max=100%) of correct answers, respectively.

For the main screen, there was one user-adversary pair that both had 100% of correct answers. They had to place five items, one of which was a big widget (size: 4×3). Both of them positioned the widget at the top left corner and used the other fields for the remaining apps and folders. Another user showed particularly bad performance and explained that although it was easy to recall the shape in which the apps, widgets and folders were placed, it was very difficult to reconstruct their exact order.

Similar observations were made for the secondary screen. There was one user-adversary pair that had 100% of correct answers. They had to place a widget that filled the complete screen so that only one placement was possible (i.e., the field at the top left corner). Another two users showed a bad performance, mostly due to the number of items they had to place (i.e., 12 apps and folders).

Number of Correct Answers (Widget Space) On average 7 out of 16 (min=4; max=10) fields of the main screen were used for widgets. For the secondary screen, the average number of fields was 10 out of 16 (min=4; max=16).

A correct answer was counted when an occupied field was marked or when an empty field was left unmarked. Table 6.5 (columns 3-4) shows the percentage of correct answers. For the main screen users had 99% (min=93.8%; max=100%) of correct answers. The number was 88.5%(min=50; max=100) for the secondary screen. In turn, adversaries had 66.7% (min=37.5%; max=100%) of correct answers for the main screen and 27% (min=25; max=100) for the secondary screen.

For the main screen, there was one user-adversary pair that both had 100% of correct answers. The corresponding adversary selected only the first two rows of the grid which is a typical position for widgets (e.g., clock).

For the secondary screen, one user had a very low number of correct answers. The reason was a dynamic widget that adapted its size. This, however, was not mirrored in the widget's internal representation (i.e., the launcher only stores the maximal possible size). Although the user provided an answer that corresponded to the widget's actual appearance, the answer did not match with the internal data and was considered as incorrect answer.

	Puzzle Tiles		Widget Space		App Selection	
	Main	Sec. Screen	Main	Sec. Screen	Main	Sec. Screen
Users	82.1%	65.2%	99%	88.5%	95%	91.3%
Adversaries	43.6%	15.2%	66.7%	61.5%	78.8%	78.8%
Difference	38.5%	50.0%	32.3%	27%	16.2%	12.5%

Table 6.5: Overview of the number of correct answers (in %) for each approach and the two screen types.

Number of Correct Answers (App Selection) A correct answer was counted when users marked apps that were correctly positioned or when they left library apps/incorrectly positioned apps unmarked.

Users had 95% (min=87.5%; max=100%) of correct answers for the main screen and 91.3% (min=75%; max=100%) of correct answers for the secondary screen. In turn, the number of correct answers of adversaries was 78.8% for the main screen (min=50%; max=93.8%) and also 78.8% for the secondary screen (min=62.5%; max=93.8%).

For the main screen, most users performed better than their adversaries. In one case, both of them had the same number of correct answers. The reason for this was that almost all apps on the grid were library apps. Since the corresponding adversary made only few selections, most of the library apps were left unmarked, resulting in a high number of correct answers.

Similar observations were made for the secondary screen. In two cases, users and adversaries had the same number of correct answers. One of these adversaries was the same adversary that made only few selections for the main screen. The same strategy was used for the secondary screen and resulted in a high number of correct answers.

When having a closer look at the incorrect answers of users, it was interesting to observe that users never gave an incorrect answer for library apps, suggesting that they are good in distinguishing between the apps that they do and do not own.

Self-Assessment and Actual Performance Users were asked to state, on a 5-point Likert scale (1=not confident at all; 5= very confident), how confident they were about the answers they had submitted. An overview of the ratings is shown in Figure 6.5 (top).

For *Puzzle Tiles* the ratings varied for the main screen. Some users were confident about their answers, while others were not. Most of them had a good assessment about their actual performance. Similar observations were made for the secondary screen.

With respect to *Widget Space*, the majority of users were confident about their submitted answers for the main screen. This is in line with their actual performance. In turn, the confidence ratings varied for the secondary screen. In most of the cases these ratings complied with their self assessment.

6 Fallback Authentication Based on Dynamic Enrollments

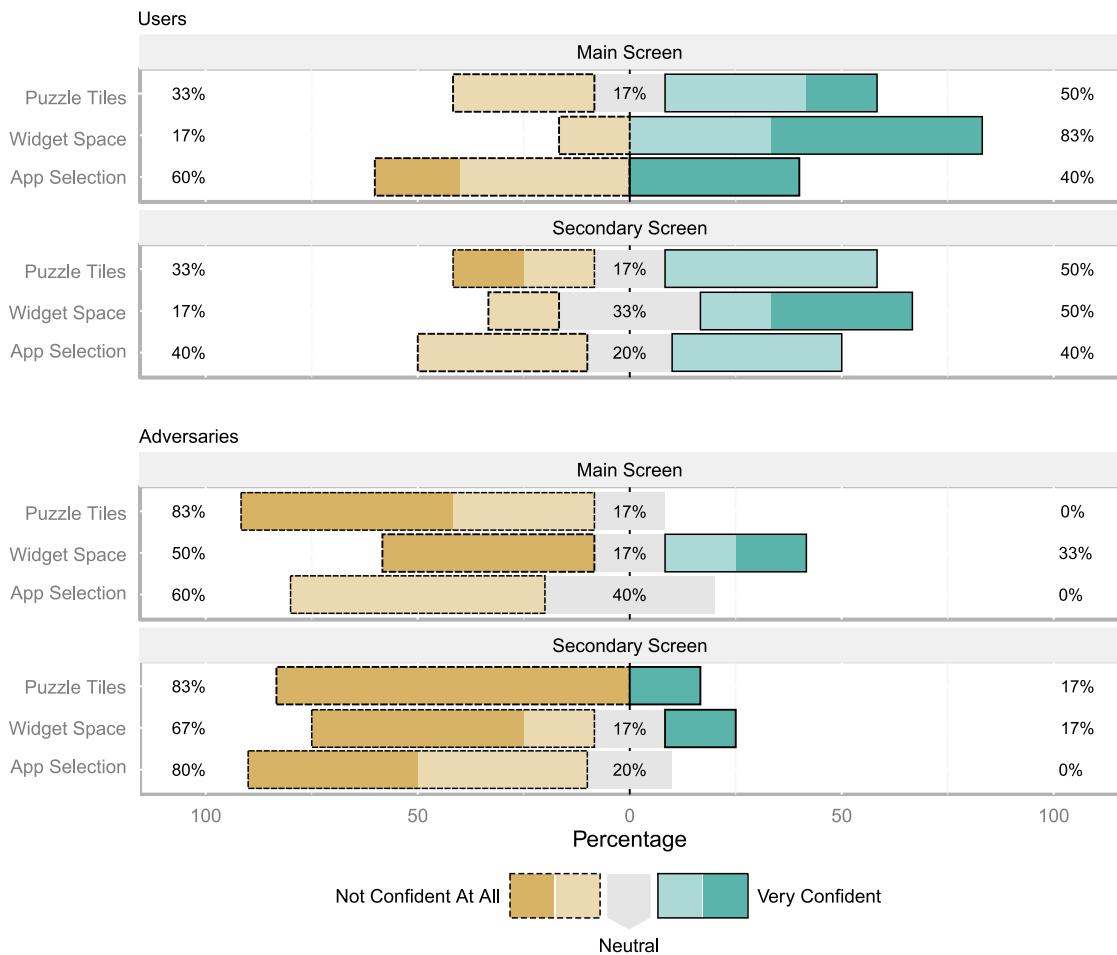


Figure 6.5: Confidence ratings by users and adversaries about their submitted answers for the three approaches and the different screen types.

The confidence ratings varied for *App Selection*. In particular unconfident users underestimated their performance. Similar observations were made for the secondary screen.

Perceived Guessability Adversaries were also asked to state on a 5-point Likert scale (1=not confident at all; 5= very confident) how confident they were about their submitted answers (Figure 6.5, bottom).

For *Puzzle Tiles*, the majority of adversaries was not confident. This applied to the main screen and the secondary screen. Most adversaries had a good self-assessment of their performance for the secondary screen, but tended to underestimate it for the main screen.

For *Widget Space*, the confidence ratings varied for the main screen. Furthermore, the assessment of adversaries was not in line with their actual performance. For the secondary screen, the majority of adversaries was not confident in the provided answers and tended to underestimate their performance.

The majority of adversaries was not confident in their answers for *App Selection*. This was the case for the main screen as well as the secondary screen. In both cases, adversaries tended to underestimate their performance.

6.2.5 Follow-Up Study

The preceding evaluation allowed us to get initial insights into the potential problems of each approach and their suitability for fallback authentication. Although *Widget Space* showed promising results, it also revealed some disadvantages. For example, there was a discrepancy between the internal representation of dynamic widgets and their actual appearance. Furthermore, there were typical widget placement patterns that, when learned by adversaries, facilitated guessing and thus are serious security threats. Due to the discovered problems that are difficult to address by further improvements (i.e., it is not possible to change the internal representation), *Widget Space* was excluded from further evaluation. Instead, the main focus of the follow-up study was to improve the other two approaches. and to evaluate them more thoroughly in terms of memorability and security.

The goal of the follow-up study was to increase the differences in performance between users and adversaries to achieve better security. Therefore, we added app icons from the app library to the app drawer of *Puzzle Tiles* to obfuscate the actual home screen items. This way, another burden is added for adversaries as they do not only have to place apps on the home screen, but also have to identify which apps belong on the corresponding home screen. Since the previous results suggested that users are good in distinguishing between the apps that they have and that they do not have, we assume that these changes will not have a negative impact on the performance of users, but instead, will make it more difficult for adversaries to guess the correct answers (see Section 6.2.4).

In the previous study, most errors for *App Selection* were made with respect to home screen apps. Thus, we inverted the authentication task. Instead of marking apps that are part of their home screens, users now have to mark the ones that are not part of it.

In the remainder of this section we report the design and evaluation of the follow-up study.

Study Design and Study Procedure

The same study design was used as for the previous study. However, instead of the screen with the most apps, the screens left and right to the main screen were used for evaluation. These screens are used more frequently [73] and thus likely to overcome the problems from the previous study (e.g., when users had to place items on less frequently used screens). The study procedure remained the same as for the previous study.

Participants

Altogether, 24 users and 24 adversaries took part in the study. All of them owned an Android device with a 4×4 home screen layout (dock excluded). Participants were recruited, using bulletin boards, social media and personal communication.

General Demographics Twelve users (6 female) and 12 adversaries (5 female) tested *Puzzle Tiles*. Users were between 20 and 29 years old (average: 24 years), while adversaries were between 20 and 31 years old (average: 25 years). Users came from different backgrounds, ranging from media, design, communication to economics and computer science. The background of adversaries included the domains health care, teaching, finance, business or computer science.

App Selection was also evaluated with 12 users (5 female) and 12 adversaries (7 female). Users were between 22 and 31 years old (average: 26 years), while adversaries were between 20 and 27 years old (average: 24 years). Users had various backgrounds, such as law, business, teaching, engineering or computer science. Adversaries had backgrounds that included teaching, law, psychology, business or computer science.

Relationship For *Puzzle Tiles*, six users brought their significant other. Another six brought a friend. There were no contradictions in the named relationships. When rating the closeness of their relationship on a 5-point Likert scale (1=not close at all; 5=very close), there was a good agreement between the user-adversary pairs. The majority of them stated to be close or even very close to each other.

With respect to *App Selection*, six users brought their significant other, four brought a family member and two brought a friend. Overall, there was a good agreement between the user-adversary pairs on the closeness of their relationship. The majority described it as close or very close.

Home Screen Organization When asked if they organize their home screens, 19 out of 24 users answered affirmatively. On a 5-point Likert scale (1=very seldom; 5=very often), 17 users stated to do this often or very often, while two users stated to do this seldom or very seldom. Users also reported that they apply different strategies for home screen organization, such as the arrangement by frequency, importance or category.

Device Sharing Altogether, 16 out of 24 adversaries stated to have used their victim's device before. Interestingly, sharing was not necessarily limited to a short time span. According to the self-report by adversaries, sharing ranged from 5-1000 minutes (average: 107.4 minutes). Thousand minutes was reported by one adversary who explained that he has access to his girlfriend's phone all the time to help her setting up the device. Other reasons for using the victim's device were playing games, looking up information (e.g., time), making calls or taking/viewing photos.

Results

Number of Home Screen Items On average, users had 7 (min=2; max=16) home screen items (i.e., icons, widgets or folders) on their main screen. For the left and right screens the average number was 6 (min=0; max=16).

Number of Correct Answers (Puzzle Tiles) A correct answer was counted when home screen items were correctly positioned; or when apps from the library were left in the app drawer. An overview of the number of correct answers is given in Table 6.6.

For the main screen, users had 68.2% of correct answers, while adversaries had 57.3% correct answers. For the left screen, users identified 69.3% correctly. The number of correct answers by adversaries was 54.7%. With respect to the right screen, users gave 74% of correct answers. In turn, adversaries answered 56.3% correctly.

Number of Correct Answers (App Selection) An answer was considered as correct, when apps from the library were marked; or when apps from the original home screen of the users were left unmarked. Table 6.6 overviews the number of correct answers.

For the main screen, users had 97.4% of correct answers. In turn, adversaries had 82.3% of correct answers. For the left screen, the number of correct answers by users was 89.6%. Adversaries answered 72.9% correctly. With respect to the right screen, 91.7% of correct answers were given by users. The percentage of correct answers for adversaries was 67.2%.

Number of Correct Answers (Comparison) A two-way ANOVA was conducted to examine the effect of *approach type* (i.e., *Puzzle Tiles* and *App Selection*) and *participant type* (i.e., user and adversary) on the overall number of correct answers. Users were significantly better in their performance than adversaries ($F(1, 44) = 15.6$; $p < 0.01$). We also found that the performance was significantly better for *App Selection* than for *Puzzle Tiles*. No interaction effects were found for *approach type* and *participant type*.

For *Puzzle Tiles*, we ran a mixed ANOVA with the between-factor *participant type* (i.e., user and adversary) and the within factors *app type* (i.e., home screen item and library app) and *screen type* (main, left and right). We found interaction effects between all three factors ($F(2, 44) = 6.79$; $p < 0.01$) and thus performed simple main effect analyses. Since this required 14 pairwise comparisons, the significance level was set to $0.05/14 = 0.004$. There was no significant effect of *screen type* on the performance of users and adversaries. However, there was a significant effect of *app type* on the performance of users ($F(1, 22) = 118.8$; $p < 0.004$). They were significantly better than adversaries for home screen apps ($F(1, 22) = 12.95$; $p < 0.004$). No significant differences were found for library apps.

Another mixed ANOVA was done for *App Selection*. Users were significantly better than their corresponding adversaries ($F(1, 22) = 10.19$; $p < 0.01$). We also found that users

		Main Screen					
		Puzzle Tiles			App Selection		
		HS	LA	Total	HS	LA	Total
User	Adversary	41.8%	95.7%	68.2%	97%	96.2%	97.4%
	Adversary	9.18%	90.4%	57.3%	76.7%	83.3 %	82.3%
Difference		32.62%	5.3%	10.9%	20.3%	12.9%	15.1%

		Left Screen					
		Puzzle Tiles			App Selection		
		HS	LA	Total	HS	LA	Total
User	Adversary	36.1%	89.2%	69.3%	58.8%	96.2%	89.6%
	Adversary	5.6%	90.8%	54.7%	53.8%	77.8%	72.9%
Difference		30.5%	1.6%	14.6%	5%	18.4%	16.7%

		Right Screen					
		Puzzle Tiles			App Selection		
		HS	LA	Total	HS	LA	Total
User	Adversary	36.4%	93.7%	74%	84.4%	92.6%	91.7%
	Adversary	7.6%	81.7%	56.3%	41.7%	73.1%	67.2%
Difference		28.8%	12%	17.7%	42.7%	19.5%	24.5%

Table 6.6: Overview of the percentage of correct answers by users and adversaries for the two approaches and the different screen types. The table distinguishes between the overall performance as well as the performance for home screen items (HS) and library apps (LA), respectively.

and adversaries were significantly better in identifying library apps than home screen apps. However, no interaction effects were found.

Self-Assessment and Actual Performance Users were asked to state, on a 5-point Likert scale (1=not confident at all; 5=very confident), how confident they were about the provided answers. In addition to this, they also had to estimate the number of correct answers they thought to have given. Figure 6.6 (top) gives an overview of the confidence ratings by users.

For *Puzzle Tiles*, the ratings varied for the different screen types. While the majority was confident in their answers for the main screen, most users were less confident for the left and right screens. While users tended to overestimate their performance for the main screen (main screen: 8; left screen: 4; right screen: 1), the majority underestimated it for the left and right screens (main screen: 4; left screen: 6; right screen: 9). Only few users made an exact assessment (main screen: 0; left screen: 2; right screen: 2). The average distance between actual performance and estimation was 20.9% (min=1.3%; max=56.3%) for the main screen, 25.3% (min=0%; max=67.5%) for the left screen and 20.6% (min=0%; max=67.5%) for the right screen.

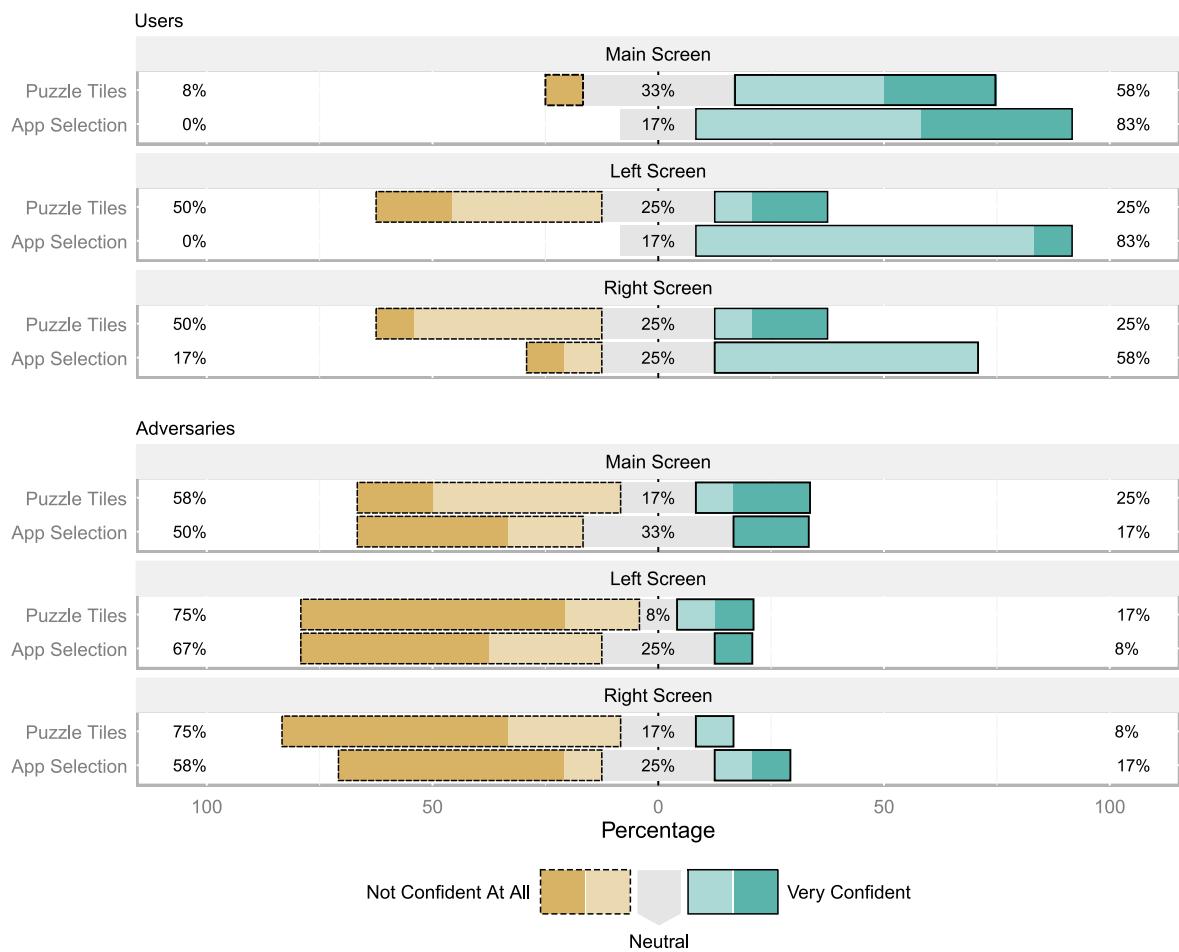


Figure 6.6: Confidence ratings by users and adversaries about their submitted answers for the two approaches and the three screen types.

For *App Selection*, the majority of users was confident in the answers they had submitted. This applies for all three screen types. Most users underestimated their performance (main screen: 7; left screen: 6; right screen: 10), but there were also users that overestimated it (main screen: 2; left screen: 2; right screen: 5) or that made an exact assessment (main screen: 3; left screen: 1; right screen: 1). The average distance between actual performance and estimation was 9.5% (min=0%; max=37.5%) for the main screen, 12.9% (min=0%; max=45%) for the left screen and 10.5% (min=0%; max=50%) for the right screen.

Perceived Guessability Adversaries were also asked to make confidence ratings and to make estimations about their performance. Figure 6.6 (bottom) overviews their ratings.

For *Puzzle Tiles*, the majority of adversaries was not confident in the answers they had submitted. This observations can be made for all three screen types. The majority of adversaries underestimated their performance (main screen: 9; left screen: 9; right screen: 10) and only few overestimated it (main screen: 3; left screen: 3; right screen: 2). The average dis-

Threshold	Puzzle Tiles						App Selection			
	Main		Left		Right		Main		Left	
	10	11	15	16	11	12	15	16	15	16
TP	9	7	4	3	8	11	8	8	8	7
TN	7	9	11	12	9	7	10	10	10	12
FP	5	3	1	0	3	5	2	2	2	0
FN	3	5	8	9	4	1	4	4	4	5
Accuracy	66.7%		62.5%		70.8%		75%		75%	

Table 6.7: The best accuracy values for the two approaches and the different screen types. The table overviews the number of required answers (i.e., threshold) as well as the number of TP, TN, FP, FN and the corresponding accuracy values.

tance between actual performance and estimation was 34.2% (min=7.5%; max=73.8%) for the main screen, 44.4% (min=2.5%; max=93.8%) for the left screen and 43.1% (min=6.3%; max=82.5%) for the right screen.

The majority of adversaries was not confident about the answers they submitted for *App Selection*. This observation was made for all screen types. Most adversaries tended to underestimate their performance (main screen: 11; left screen: 11; right screen: 10) and only few of them overestimated it (main screen: 1; left screen: 1; right screen: 2). The average distance between actual performance and estimation was 39.4% (min=6.3%; max=75%) for the main screen, 40.6% (min=11.3%; max=77.5%) for the left screen and 38.9% (min=3.8%; max=75%) for the right screen.

Accuracy So far, we have reported the number of correct answers by users and adversaries, respectively. However, we did not interpret them with respect to an actual authentication system. Similar to the previous studies, accuracy calculations were used to identify the number of required answers in order to authenticate successfully. Since the maximum possible number of correct answers is 16, we used this value as an upper threshold to calculate the accuracy values, true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) when 0 to 16 correct answers were required.

For *Puzzle Tiles*, the highest accuracy values ranged between 66.7% and 70.8% and required users to have between 11 and 16 correct answers. The highest accuracy values for *App Selection* ranged between 75% and 79.2% and required between 15 and 16 correct answers. The number of FP and FN for the corresponding accuracy values are listed in Table 6.7.

6.2.6 Discussion

The accuracy results for *Puzzle Tiles* and *App Selection* are unacceptable for an authentication scheme, meaning that both approaches, as is, are not suitable for fallback authentication. Nonetheless, the results are valuable to teach us the key aspects that should be considered whenever icon arrangements on smartphones are exploited as a data source.

Data Availability

The number of home screen items varied from user to user. While some of them had only few items on the tested home screens, others had many. Users with few items often placed them on the main screen, leaving the left and right screens empty. This means that not all screens can be used for authentication when these users are locked out. However, testing only one screen is likely to be insufficient to reach an acceptable level of security. Thus, it is advisable to use icon arrangements only for users that have enough data. In case there is not, users must be provided with an alternative for which enough data is available.

Our evaluation focused on screens that were shown to be used the most frequent [73]. Users may have more apps on other screens. However, the first study revealed that these screens are sometimes used as trash spaces for unused apps. This makes recalling their exact arrangements challenging. In order to increase data availability, one could think of detaching home screen apps from their actual location. For example, instead of asking where an app is positioned, the task could be simplified by asking whether the given app exists on one of the home screens.

User-Adversary Performance

Users were significantly better than their adversaries. Nonetheless, the performance distance between them was too small to reach a satisfying security level (i.e., to distinguish between users and adversaries during authentication). Interestingly, the confidence ratings showed that, most of the time, adversaries did not know, but guessed the correct answers.

This suggests that the answer space of icon arrangements is not large enough, making it easier for adversaries to make correct guesses. One of the key reasons for this is certainly the position and size of widgets that limit the answer options. We tried to overcome this problem by obfuscating the presence of widgets with the use of library apps. However, based on the results, we recommend to filter widgets from the arrangement task in order to increase the performance gap between users and adversaries.

Another reason for the removal of widgets is the existence of dynamic widgets that adapt their size to the content they hold. Since the internal representation only includes the maximum possible size, dynamic widgets are a potential source of error when the submitted answers are evaluated. In the first place, it would be desirable to include the current size of widgets in the launcher's database. However, there is no common standard to enforce launcher developers to include this information. It is recommendable, to filter widgets from the authentication task and to focus on app icons and folders instead.

When comparing the results from the first study and the follow-up study, the performance of users, in some cases, seem to be worse for the latter. For example, for *Puzzle Tiles* users gave 82.1% of correct answers for the main screen, while the percentage was only 68.2% in the follow-up study. One reason could have been that there were less tech-savvy users in the follow-up study who did not arrange their home screen items on their own (see Section 6.2.6).

Home Screen Items and Library Apps

Users were good in identifying the apps that they have and do not have on their home screens. However, with respect to the arrangement of apps, most users reported that, although they had a vague idea of the app positions, they had difficulties to recall the exact location. This means that the factor position should be neglected, simplifying the idea of app arrangements to their mere presence.

We further found for *App Selection* that adversaries were better in identifying library apps than home screen items. This suggests that the used app library needs to be optimized to make it more difficult for adversaries to distinguish between the two app types.

User-Adversary Relationships

After the study, some user-adversary pairs gave us further feedback about the study task. In some cases, the adversary explained to have set up the phone for the corresponding user and thus was familiar with the user's icon arrangements. In turn, the corresponding user reported difficulties in recalling the arrangement as it was not done by herself.

Although, these kinds of adversaries are likely to know the correct answers, it can be assumed that they are not a serious threat (as they already have access to the phone to set up things). On the contrary, they are probably the user's first contact person in case of lockout to help them to regain access. Nonetheless, the use of icon arrangements for fallback authentication is cumbersome for these users (as they have to rely on a third person). Therefore, other types of fallback schemes would be a better fit for them.

6.3 Installed Apps

Although the use of icon arrangements for fallback authentication showed many issues, the problems were a good indicator for the needed improvements for our next design. The main goal was another app-related approach that a) is independent of app positions, b) has a better data availability, and c) increases the performance gap between users and adversaries.

Based on these requirements, we came up with an approach based on installed apps. The basic idea is to ask users whether an app is installed on their device. This has the advantage that the position factor can be neglected. Furthermore, we are not limited to the users' home screen items, but can extend the data source to all installed apps to address the problem of data availability. The question whether this approach improves the gap between users and adversaries is the main objective of the study.

In the remainder of this section, we provide further details about the described approach. We provide a description of the study prototype, the study design and study evaluation. The results are reported and discussed at the end of this section.

6.3.1 Approach

In our approach, users are consecutively shown app icons and their names. The apps can either be installed on the user's device or taken from an app library. The library apps are needed to act as distraction items. For each shown app, users have to decide whether it is or is not installed on their devices. In case they identify a certain number of apps correctly, they are authenticated. Otherwise, their authentication attempt is denied.

The main assumption is that users are able to distinguish between the apps that they have and do not have (as also shown in the previous study). This assumption is further supported by work from Sun et al. who tested a similar app-related approach with a success rate of around 95% [167]. Although the chances are high that adversaries identify certain apps correctly, we assume that it will be difficult for them to have a sufficient number of correct answers for successful authentication.

6.3.2 Prototype

The prototype for the user study was implemented for Android smartphones with Jelly Bean (version 4.1) or higher. The three main components of the app are described in the following.

Device Apps

In the first step, the study application scanned the storage of the user's device for all installed packages to generate a list of installed apps. In the second step, we removed all apps from the list that were marked with a system flag. This included pre-installed apps and system-related functionalities (e.g., download manager or installed certificates). Pre-installed apps were removed because they are, as their name implies, installed without the knowledge of the user and thus could have negatively influenced the identification of (not) installed apps. System-related functionalities were excluded to avoid overwhelming users during authentication as there are usually hundreds of system apps installed on a device.

Library Apps

The list of device apps was complemented with apps that are not installed on the device (i.e., library apps). Since we wanted to create a diverse library to prevent apps from being too similar (and maybe easier to guess), the apps were selected from different Google Play Store categories (as of May 2014). Altogether, the library consisted of 49 apps. This included 15 paid and 15 unpaid apps from the list of top apps as well as a random selection of 19 less popular apps from different app categories (e.g., sports or education).

Question Application

For the actual study task, users were asked about all device apps and a random number r of library apps, with r ranging from 0 to the maximum number of apps in the library (after

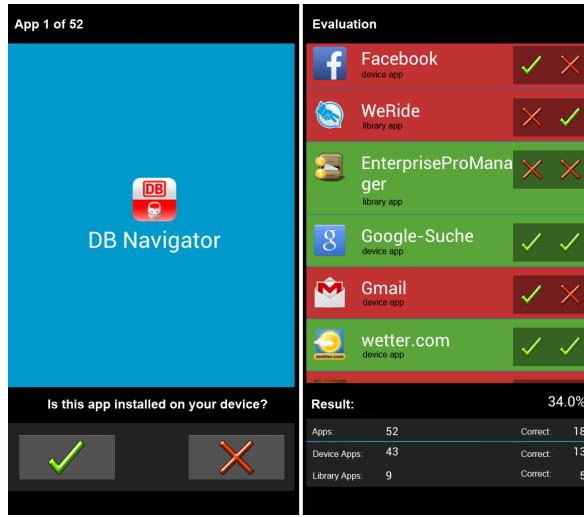


Figure 6.7: Screenshot of the study application. On the left: an example question for an app with the two answer options *yes* and *no*. On the right: overview of the performance of a user.

removing apps that are actually installed on the device). The decision to quiz all installed apps was done for evaluation purposes, for example, to see how many installed apps users recognize as such. In a real-world setting, one would use only a subset of those apps for authentication.

For each app in the quiz, we asked users whether it was installed on their device. The question was accompanied by the corresponding app icon as well as app name. To answer the questions, users could choose between *yes* or *no* (Figure 6.7, left). Although the chances to guess the correct answer from two options is 0.5, we have to keep in mind that this probability decreases with each additional question.

Users did not receive immediate feedback about the correctness of their answers. However, we showed them an overview of their performance once they had finished the study task (i.e., after they had answered all questions; Figure 6.7, right). This was done to encourage discussions about the apps, for example, to reveal different answering strategies. In a real-world setting, such an overview must not be shown to prevent adversaries from learning the correct answers for future attacks.

6.3.3 Threat Model

Similar to the preceding threat models in this chapter, we assume an adversary close to the user and with advanced knowledge about the user (see Section 4.3.2).

The chance for a random adversary (when equal answer distribution is assumed) to guess the correct answers is $(\frac{1}{2})^x$, with x depicting the number of apps asked about. However, the chance to guess the correct answers is likely to depend on the popularity of the apps that

the adversary is asked about. In this case, the chances should include a weighting factor: $\prod_{i=0}^x \frac{w_i}{2}$. For example, the weight could be calculated using the percentage u of Android users who have a certain app installed. If more than 50% of users have app i installed, then the weighting factor should be $w_i = 1 + (\frac{u}{100})$ (i.e., adversaries are more likely to answer *yes*). In turn, if less than 50% of users have app i installed, the weighting factor should be $w_i = 1 + (\frac{100-u}{100})$ (i.e., adversaries are more likely to answer *no*). In all other cases, the weighting factor should be 1.

6.3.4 User Study

Study Design

The user study consisted of one task (i.e., identifying apps). This task had to be completed by all participants and we used the percentage of correct answers as a measure for performance.

Participants were recruited over bulletin boards, social networks and personal communication. Similar to the previous designs in this chapter, we had two types of participants: users and adversaries. We will continue using this wording to distinguish between them. In order to take part in the study, users were required to own a smartphone. They were also asked to bring another person they are close to who then acted as adversary. During recruitment, we gave them examples for close persons (e.g., significant others or best friends). Although participants were informed about the general topic of the study (i.e., fallback authentication), we did not disclose any additional information about our app-related approach. This was done to prevent users and adversaries from looking up their apps in advance.

Study Procedure

Participants were invited to our lab for the study. After giving them a brief introduction about fallback authentication, adversaries had to leave the room and wait outside. Then, we asked for permission to install the study application on the users' device. Once installation was done, the study began and users had to complete the study task (i.e., marking apps as installed or not installed). This was concluded with a demographic questionnaire and a brief post-task interview. During the interview, we went through the list of apps that users were asked about and encouraged them to discuss their answer strategies and problems.

The same procedure was repeated for adversaries. Altogether, the study lasted for about an hour and users and adversaries received 10 € gift vouchers each for their participation.

Participants

General Demographics Altogether, 15 users (5 female) and 15 adversaries (12 female) took part in the lab study. While users were between 20 and 34 years old (average: 24 years), adversaries were between 21 and 49 years old (average: 26 years).

	Device	Library	Total
Users	95.4%	95%	95.2%
Adversaries	60.5%	82%	68.9 %

Table 6.8: Overview of the number of correctly identified apps for users and adversaries (in %).

Most users were students with a technical background (e.g., computer science, engineering, etc.). Five participants were employed in different domains (e.g., sales, agriculture, etc.) and one participant was a high school student. Similar demographics were found for the adversaries. Nine of them were students, mostly with technical backgrounds (e.g., computer science, physics, etc.). Five of them were employed (e.g., sales, administration, etc.) and one adversary was a homemaker.

Relationship The relationships between users and adversaries were diverse. Seven of them brought their significant other, seven brought their close friend and one brought a family member. When asked to rate the closeness of their relationship on a 5-point Likert scale (1=not close at all; 5=very close), most users described it as very close or close. Only in one case was the relationship considered neutral. Most of the time, these ratings were in line with the ratings by the corresponding adversaries. However, there were three cases in which the perceived closeness was not mutual. For example, while one user found the relationship as very close, the corresponding adversary described it as neutral only. We did not disclose these discrepancies to our participants.

Device Sharing The majority of users (11 out of 15) stated that they had shared their device with the corresponding adversary before. This was done, for example, to show photos or to make calls. However, they also noted that device sharing rarely happened (between 1 and 4 times) and was most of the time limited to short time spans (between 2 and 30 minutes). This information is interesting to see if adversaries had had the chance to spy on the user's apps before they participated in the study.

Results

Number of Installed Apps The number of installed apps on a device varied from user to user. There were users with very few apps (< 20), but there were also users that possessed a lot of them (> 70). On average, users had 43 apps (min=14; max=91) installed on their devices. Interestingly, users tended to underestimate the number of apps they had. On average, the difference between estimation and actual number of apps was 25 (min=2; max=82).

With respect to their installation behavior, users stated that they install on average two apps per month (min=1; max=5). Most of them preferred free apps and only four of them stated to be willing to pay between 1€ and 5€. In turn, the frequency in which these apps were uninstalled varied from user to user. Deleting apps was done occasionally by six users, often by three users, seldom by four users and never by two users.

Number of Correct Answers On average, users had to identify about 72 apps (min=40; max=104), while the average number was 70 apps (min=37; max=109) for adversaries. Table 6.8 overviews the percentage of correct answers that were given by users and adversaries, respectively. Altogether, users were better in identifying apps (95.2% correct answers) than adversaries (68.9% correct answers).

While the performance of users was about the same for device apps and library apps, respectively, the results suggest that adversaries were better in identifying device apps than library apps. Therefore, a mixed ANOVA with the between-groups factor *user type* (i.e., user and adversary) and the within-factor *app type* (i.e., device app and library app) was conducted. We found that users were significantly better in identifying apps than their corresponding adversaries ($F(1, 28) = 66.4, p < 0.01$). However, no main effects were found with respect to *app type*. There were also no interaction effects.

Authentication Time Time measurement started when the first question of the quiz appeared and ended with the submission of the answer to the last question. On average, users needed 172.5s (min=96.1s; max=280.6s) to complete the study task, while adversaries took about 444s (min=242.6s; max=769.3s).

Since the number of questions to be answered was not the same for all users, we also measured the time that was needed to answer each individual question. On average, users needed 2.4s (min=1.8s; max=3.3s) to mark an app as (not) installed. In turn, adversaries needed on average 6.5s (min=5s; max=8.6s).

A mixed ANOVA with the between-groups factor *user type* (i.e., user and adversary) and the within-factor *app type* (i.e., device app and library app) was conducted. Hereby, the average time per user was used for the analysis. The results showed that users were significantly faster than their corresponding adversaries in identifying apps ($F(1, 28) = 154.9, p < 0.01$). No main effects were found for the factor *app type*. There were no interaction effects.

Self-Assessment and Actual Performance Before revealing their actual performance, we asked users and adversaries to estimate the percentage of correct answers they thought to have given. The average estimation by users was 89.1% (min=75%; max=100%) and was 56.6% (min=10%, max=99%) for adversaries. The adversary who thought to have identified 99% of apps correctly, had only 68.5% of correct answers in the actual performance. The average difference between estimated and actual performance was 7% (min=0; max=25%) for users and 25% (min=0.4%; max=71.7%) for adversaries.

Device Familiarity and Performance In order to see if the familiarity of adversaries with the user's device influenced their performance, we asked them to provide familiarity ratings using a 5-point Likert scale (1=not familiar at all; 5=very familiar). Six adversaries were completely unfamiliar with the user's device. Five adversaries were (very) familiar with the device. The remaining ones had a neutral opinion about their familiarity.

6 Fallback Authentication Based on Dynamic Enrollments

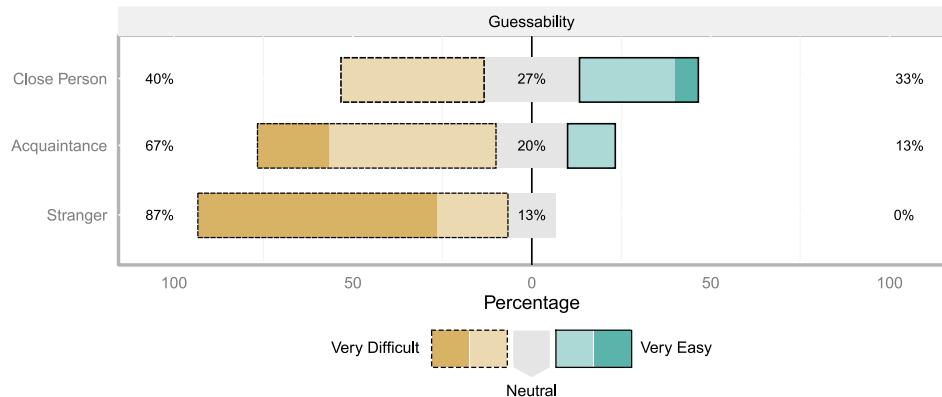


Figure 6.8: Overview of the assumed guessability of installed apps by users for different types of adversaries (i.e., close persons, acquaintances and strangers).

A Spearman's rank-order test was conducted to analyze the correlation between familiarity and performance. However, we did not find any statistically significant correlations.

Perceived Guessability In the study questionnaire, users were asked to rate how easy they think it is for different types of adversaries (i.e., close persons, acquaintances and strangers) to identify their apps correctly. The ratings were given on a 5-point Likert scale (1=very difficult; 5=very easy). While the majority of users believed that it is difficult for strangers (13 users) and acquaintances (10 users) to guess the correct answers, the opinions were ambiguous for close adversaries. Six users thought that it is difficult for close persons to identify the apps correctly, while five users stated the opposite opinion. An overview of the given ratings for all three types of adversaries is given in Figure 6.8.

Guessing Strategies During the post-task interviews, we asked adversaries about their guessing strategies, one of which was related to the popularity of apps. Social apps (e.g., Facebook) or communication apps (e.g., WhatsApp) were often assumed to be widely-used and lead to the decision to mark them as installed.

Other strategies took the user's characteristics into account. This included the user's geographic location, personal preferences or financial situation. For example, one adversary told us that she marked all apps as installed that fit into her boyfriend's special taste in games. Other adversaries assumed their victim to be too cheap to pay for apps and thus identified all apps that were labeled with *premium*, *plus* or *full version* as not installed on the device.

Preceding selections were also an influencing factor in decision making. When adversaries had marked an app from a specific category as installed, they were hesitant to do so again in case a similar app was shown (e.g., multiple weather apps).

User Problems Similar to the interviews with adversaries, we encouraged users to tell us about their problems during decision making. Most users had difficulties in distinguishing between similar apps. For example, there are numerous versions of the puzzle game 2048.

It was difficult for users to decide if the version they were asked about was the one they had installed on their device. They also reported difficulties when they had to decide whether an infrequently used app was still installed on their device or not.

Accuracy With respect to the overall performance, users were always better than their corresponding adversaries in identifying apps. Furthermore, the worst user was better than the best adversary (89% versus 81.7%). If we had asked users about all their apps for authentication using a threshold of 82%, all users would have been authenticated successfully, while all adversaries would have been rejected. This naive observation is promising. Therefore, we move forward to a more formal evaluation using the accuracy metric (see Section 4.3.3).

In theory, it would be necessary to consider all possible combinations of the apps that users and adversaries were asked about (subsets included). With 15 users and 15 adversaries, the number of possible combinations is:

$$\sum_{i=1}^{15} \left[\left(\sum_{k=1}^{u_{i_max}} (u_{i_max})_k \right) + \left(\sum_{j=1}^{a_{i_max}} (a_{i_max})_j \right) \right],$$

with u_{i_max} and a_{i_max} representing the maximum number of apps in the quiz for the i^{th} user and adversary, respectively.

However, due to computation limitations, we had to follow another approach in practice. That is, for each user/adversary, we randomly selected x apps from the list of apps, with x ranging from 15 to 45. The upper border was based on the average number of apps that users had installed on their devices and was chosen to maintain a certain level of data availability and comparability of the results (e.g., not all users had enough apps to be asked about more than 45 apps).

This procedure was repeated a thousand times, resulting in 15000 ($= 15 \times 1000$) authentication attempts by users and 15000 ($= 15 \times 1000$) attacks by adversaries with different combinations of x apps. Please note that for $x = 40$, we only had 14000 attacks. For $x = 45$, only 14000 authentication attempts and attacks were taken into account. The reason was that some users/adversaries were asked about fewer apps during the study.

Based on this data, the number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN) and accuracy values were calculated. The calculation took into account the number of asked apps in the quiz and the number of allowed errors e , with e ranging from 0 to 10. Table 6.9 overviews the results of the analysis. The best accuracy values were yielded when 40/45 apps were used. Further details are provided in the following.

Asking users to identify 40 apps and allowing at most 5 errors, resulted in an accuracy of 94.9%, (747 FP; 722 FN). Users succeed 95.2% of the time. In turn, the success rate of adversaries is 5.2%. Similar observations were made when using 45 apps and 6 allowed errors as parameters for the quiz. The accuracy was 95.2% (824 FP; 529 FN). The success rate was 96.2% for users and 5.8% for adversaries.

		Number of Asked Apps						
		15	20	25	30	35	40	45
Allowed Errors	0	77.1	73.9	71.4	69.5	67.8	65.6	63.3
	1	87.8	85.5	83.1	79.9	77.2	74.5	71.3
	2	88.1	90.7	90.2	88.1	85.7	82.4	79.5
	3	82.0	89.4	92.5	92.6	91.5	89.2	86.8
	4	73.1	84.4	90.4	93.5	93.9	93.5	91.8
	5	64.7	77.0	85.8	91.1	94.0	94.9	94.5
	6	58.6	69.8	79.7	87.1	91.6	94.1	95.2
	7	54.7	63.2	73.3	81.9	87.8	91.7	94.3
	8	52.4	58.2	67.0	75.9	83.3	88.9	91.9
	9	51.1	55	61.7	70.2	78.3	84.8	88.7
	10	50.4	52.8	57.9	65.2	72.6	80.4	85.2

Table 6.9: Overview of accuracy values (in %) with respect to the number of apps asked in the quiz and the number of errors allowed. The best accuracy values are highlighted in grey.

6.3.5 Discussion

Using installed apps for fallback authentication achieved promising results. In comparison to the icon arrangements study, we were able to increase the performance gap between users and adversaries and influenced the accuracy values positively. With 95% the results are comparable to our first study about smartphone activities. However, this time, we had a false positive rate of 5%, meaning that some adversaries were able to authenticate successfully. Therefore, we identified several areas for improvement that are supposed to reduce the number of false positives for future implementations.

Data Availability

The use of installed apps as data source increased the data availability in comparison to the icon arrangements study. Nonetheless, similar observations were made with respect to the different user types: While some smartphone owners used few apps, others had many apps installed. In particular for the former type of user, there may not be enough data to create a sufficient number of questions. This problem is mitigated by the inclusion of library apps. In our implementation the number of library apps was defined by a random number between zero and the app library's size. For future implementations, it is advisable to increase the lower bound, for example, by using the difference between the required number of questions and the number of installed apps available on the device.

Similarity of Apps

When users were asked about different versions of the same app (e.g., free and premium), they had difficulties in distinguishing between them. This kind of confusion should be avoided, for example, by merging different app versions. Instead of using the terms *Weather Free* or *Weather Premium*, users should be asked about the *Weather* app only.

The similarity of apps was also exploited by adversaries as guessing strategy. They tended to mark an app as not installed in case they had already been asked about a similar app prior to this. Such behavior can be used to improve the quiz design. For example, library apps that are similar to the installed apps of the user should be asked about first. This way, adversaries are likely to mark installed apps as not installed. However, this modification must be done in such a way that the randomness of the quiz is not hurt.

Popularity of Apps

In other guessing strategies, adversaries assumed that popular apps (e.g., Facebook) were installed on the victim's device and marked them accordingly. This approach is reasonable and has different implications. Popular apps should be either removed or weighted, meaning that less popular apps contribute more to the overall authentication score. The fact that adversaries are more likely to mark popular apps as installed can also be used to mislead them. In case a popular app is not on the victim's device, it should be included in the question set. This way, adversaries are encouraged to provide an incorrect answer.

Time Limit

Users in our study were significantly faster in answering the questions than adversaries. While they needed on average 2.4s per question, adversaries needed more than twice as long. It took them on average 6.5s. This could be used to add a time limit for each question or to look at the overall time when deciding whether authentication should be considered successful. This has the advantage that there is not enough time for adversaries to apply their answering strategies. For example, it will be difficult to research answers on social networks (e.g., to see whether information was shared using a certain app).

Library Apps

Based on the results from the icon arrangements study, we created a more diverse app library that was based on various lists from the Google Play Store. Since these lists change frequently (app icons are updated, apps are removed or new apps are released), it is important for a real-world deployment to work with a dynamic library that is updated on a regular basis. For example, a library update could be done when users update their operating system. Such an update is necessary to prevent adversaries from recognizing library apps.

6.4 Lessons Learned

In this chapter, we explored the potential of different information categories on smartphones for fallback authentication with dynamic enrollments. In three main projects, various examples were implemented and evaluated. This helped us to identify promising information categories and, beyond that, taught us the challenges when handling and evaluating personal smartphone data. The following list summarizes the key lessons.

- **Trade-Off between Usability and Security.** The best solution in terms of memorability is not necessarily the best choice in terms of security. For example, although questions about communication activities had good recall rates, these questions were also easy to guess for close adversaries. In turn, the best trade-off was found for app-related questions. The decision which factor should be prioritized depends on the requirements of the system to be implemented. In most cases, a balance between them is desirable. Since fallback authentication happens infrequently, memorability is an important factor. But at the same time, fallback schemes need to be at least as secure as other common authentication schemes on smartphones (e.g., PIN) to prevent unauthorized access.
- **Privacy.** Since the use of smartphone data for fallback authentication reveals information to potential adversaries, privacy is another important factor to be considered. For example, the use of photos was not well received by our study participants due to the sensitivity of this data type. The information used must be distinct enough to authenticate users/reject adversaries successfully, but at the same time, must not be too private to be shared with others.
- **Confidence Ratings.** The extent to which users and adversaries are confident about their submitted solutions is a good indicator for the memorability and guessability of different information categories. This can be useful to explore the potential of different information categories at an early stage before the actual implementation, for example, with semi-structured interviews.
- **Data Availability.** Dynamic enrollments rely heavily on the availability of data to provide users with the needed fallback schemes. However, there are situations in which not enough data is present to generate the authentication task. For example, there were smartphone owners in our studies who install and use very few apps only. In those cases, it is important to have a fallback for the fallback scheme. In general, it is advisable to have at least one authentication scheme in the fallback chain that is based on static enrollment as an alternative option.
- **Passive and Active Consumption.** Some information categories work better than others. In particular, information that is consumed passively (e.g., music, app positions) showed bad recall rates. Instead, information categories that require the active involvement of users should be favored for dynamic enrollments due to memorability reasons.
- **Information Category: App.** Fallback authentication based on smartphone apps showed the best trade-off between the three factors memorability, security and privacy. Thus, this information category is a promising data source for dynamic enrollments in the context of fallback authentication. However, as shown by the evaluation of different examples, not all implementations worked equally well. The success of this information category depends on the way it is used for the design of fallback schemes.

-
- **Distractor Items.** Although users do not mind to reveal app-related information to others, the use of distractor items is advisable to obfuscate the apps that exist on a device and to sustain the user's privacy. The composition of distractor items must be done carefully, taking into account that apps change frequently (e.g., icons) and need to be updated on a regular basis.
 - **Guessing Strategies.** Since most users have to guess the answers to app-related questions, they apply different guessing strategies (e.g., popularity of apps). Finding out about these strategies is important to use the insights for the design of the fallback scheme (e.g., by filtering apps or weighing apps).

7

Assembling the Pieces

The main objective of this thesis was to get a better understanding of fallback authentication as well as a thorough exploration of different types of autobiographical memories for the design of fallback schemes. We did not strive for the perfect solution to be used in practice, but, instead, we were interested in the potential and challenges of using different types of autobiographical memories that go beyond personal facts.

We found that lockout experiences on smartphones are rare and that most users are satisfied with current fallback solutions. Nonetheless, we also identified special circumstances in which recovery from lockouts is inconvenient and difficult (e.g., when users are on the go). These circumstances yield for alternative fallback schemes that are independent, immediate and engaging to complement (but not replace) currently available solutions.

For the design of alternatives, our choice fell on the use of autobiographical memories for several reasons: First, the retrieval of autobiographical memories is independent from third parties as it relies on information about the self. Second, the retrieval is immediate as the required information is stored in the users' head, and third, it is engaging as users have to actively provide this information and, in addition, can get new insights about themselves.

Although not all approaches with autobiographical memories worked as we had expected, each example implementation taught us important lessons about their requirements (Section 7.1), design (Section 7.2) and evaluation (Section 7.3) in the context of lockout situations. While the discussions in the preceding chapters focused on specific prototypes, this chapter brings each single piece together to create the big picture of this thesis and to discuss their applicability for other application domains (Sections 7.4 and 7.5).

7.1 Motivation for Fallback Authentication

In order to learn about the frequencies, problems and countermeasures of mobile fallback experiences, we conducted an online survey and complementary interviews to get a better overview on this matter (Chapter 3). It was not surprising that lockouts on smartphones are rare, but the countermeasures that some people take when special circumstances apply were intriguing (e.g., disposal of phone) and thus asked for alternative solutions. Since only a small target audience is affected by these problems, it is justifiable to question the necessity of alternative fallback schemes and research in this area. However, our opinion on this issue is clear based on several reasons.

First, as usability and security researchers, we always strive for the better and are not only responsible to address the problems of the general population, but we are also responsible to help those that are not part of it. And although lockout situations are rare, they can come with extreme inconvenience so that better solutions are needed.

Second, a remarkable number of researchers design systems well applicable for fallback authentication, but are not aware of it and propose their systems for primary authentication instead (e.g., [36, 37, 174]). Although these systems are often interesting, they are less convincing due to the typical constraints of primary schemes, such as long authentication times. The framework in this thesis enables researchers to categorize their ideas based on different properties and to learn which circumstances they are designing for. This allows to define convincing use cases that highlight their relevance to push further research in that area.

7.2 Designing for Fallback Authentication

Our framework for fallback authentication provided a clear definition of fallback authentication and depicted its characteristics to overview the options for the design of fallback schemes. We implemented five different prototypes that varied in the type of scheme they used, the type of memory they exploited as well as the type of enrollment and authentication they required. From each of these implementations we uncovered different design implications that should be considered in the context of fallback authentication. The key insights are summarized next.

7.2.1 Type of Authentication Scheme

Authentication schemes can be based on something you know, something you are and something you have [187]. Since the last does not meet the requirement of immediacy, our example implementations focused on things the user knows (i.e., location-based memories and recent smartphone activities) and on things the user is (i.e., sketching skills).

	Project	Adversary Type	Time	Successful Attacks	Failed Auth.	Accuracy
Static	2-Finger Sketch	Strangers	4 weeks	32.6	37.6	67.1
	Location-Based Questions	Close Persons	5 min	0	3.3	98.3
			1 week	0	3.3	98.3
			4 weeks	0	10	95
			6 months	0	16.7	91.7
		Strangers	5 min	0	3.3	98.3
			1 week	0	3.3	98.3
			4 weeks	0	10	95
			6 months	0	16.7	91.7
Dynamic	Smartphone Activities	Close Persons	n/a	9.1	0	95
	Acquaintances	n/a	0	0	100	
	Icon Arrangements (Puzzle Tiles, Main Screen)	Close Persons	n/a	25 / 41.7	41.7 / 25	66.7
	Icon Arrangements (App Selection, Main Screen)	Close Persons	n/a	16.7	33.3	75
	Installed Apps	Close Persons	n/a	5.9	3.8	95.2

Table 7.1: Comparison of the successful attacks by adversaries, failed authentications by legit users and the overall accuracy values for the different projects in this thesis.

The results for our biometric approach were disappointing: Drawing sketches with two fingers simultaneously was hard to learn and lead to user dissatisfaction that, in turn, caused undesired behaviors. Also in terms of security, our approach was not convincing as there was a high number of false positives and false negatives (Table 7.1). Overall, the results did not encourage the use of biometric schemes (as we implemented them) for fallback authentication. All things considered, knowledge-based schemes showed the best potential for the design of fallback schemes in terms of memorability and security and should be favored over other design options.

7.2.2 Autobiographical Memories

The knowledge-based fallback schemes in this thesis exploited different autobiographical memories that included episodic memories from the very past (i.e., location-based questions) and episodic memories from recent smartphone activities. While some of these memory types showed bad usability and security properties, others worked very well (Table 7.1).

Location-based questions, for example, yielded very high recall rates and had, at the same time, very low guessing rates. The good memorability properties were due to the strong location-depended context of these questions, but also due to the way the answers provision was implemented: Selecting a location on a map (instead of text-based input) allowed

users to recall locations with the help of surrounding landmarks. In turn, these questions were hard to guess due to their personal nature, but even when the answer was known, it was difficult to make precise selections on the map that were close enough to the actual solution. With respect to the different types of smartphone activities, users found it easier to answer questions about recent activities (e.g., yesterday or last week) that were actively created by them. Overall, the best trade-off between memorability and security was found for app-related questions. But how well certain questions could be answered also depended on the way the authentication task was implemented. For example, while it was difficult for users to recall the exact position of their apps, they showed a better performance when the authentication task was simplified to yes-no options.

All things considered, one can say that the suitability of autobiographical memories for fallback authentication depends on two factors: The characteristics of the corresponding memories and the way these memories are exploited in the implementation of the challenge-response design.

Overall, the reasons for the success or failure of different memory types as enablers of fallback authentication can be summarized as follows: Some of them were not distinct enough (and thus easy to guess), lacked in active involvement of the user (and thus hard to recall) or were too privacy sensitive (and thus should not be used). In turn, other memory types were personally relevant (and thus memorable), had a strong spatio-temporal context (and thus also memorable) or were personal, but not too private (and thus should be used). As a conclusion, it is advisable to take advantage of episodic memories from the declarative knowledge that fulfill the following requirements: First, they are actively created by the user and not only the byproduct of other activities. Second, they possess a spatial or temporal context. Third, they are personal, but not too private to be shared.

7.2.3 Type of Enrollment

In the scope of this thesis, we analyzed two enrollment types: static enrollments that require users to actively provide authentication-related information before fallback authentication takes place; and dynamic enrollments that collect this information implicitly in the background during the user's interaction with the device.

With respect to static enrollments, we found that learnability is an important factor to consider for their design in order to achieve user satisfaction. For example, our approach with 2-finger sketches did not work out due to its bad learnability properties that lead to user dissatisfaction. But also the promising location-based questions left room for improvement: If users had learned earlier about the required precision of their answers, this knowledge would have positively influenced the recall rates of the system.

These kinds of problems do not exist for dynamic enrollments, but this does not spare them from other issues that must be taken into account (i.e., privacy and data availability). Since dynamic enrollments are based on implicit information, they are using content on the user's

device without their explicit consent. There is always the risk of revealing information that users would prefer to see protected. We encountered this problem, for example, when we used photos for question generation. Therefore, privacy-sensitive information should not be used for the design of fallback schemes, but since different users may consider different types of information as sensitive, it is advisable to inform them about the availability of dynamic fallback schemes and the data sources that these schemes use. This way, users can opt out data sources that they do not want to be used.

Another issue that we observed was related to data availability. There were users that did not generate enough data for certain type of questions (e.g., users that have only few apps). Therefore, it is important that dynamic fallback schemes do not rely on one data source only, but take into account multiple data sources for question generation to ensure better data availability. However, in the last resort, when all data sources are insufficient, a fallback option for the fallback is needed.

The discussed problems show that both approaches have their advantages and disadvantages (see Chapters 5 and 6), but instead of considering them as competing approaches, it is advisable to view them as complementary solutions that help balance each other's shortcomings. Overall, we recommend the following fallback chain: A primary scheme that is used on a daily basis to access the device, a dynamic fallback scheme that is used when the primary scheme fails, and a static fallback scheme that is used when the dynamic scheme is not feasible (e.g., when not enough data is available). This way, users have different options to choose from in case lockouts happen and can use the scheme that is best for their current context (e.g., home versus vacation). At the same time, the limited number of options in the fallback chain keeps the number of potential points for attacks small to maintain a reasonable level of security.

7.2.4 Type of Challenge, Type of Response

As previously described, the memorability and security of autobiographical memories is also influenced by the way the corresponding authentication task is implemented. The different options for implementation are numerous and there is no general suggestion for the perfect solution, but our observations allow to point out important aspects that should be taken into account during the challenge-response design.

Challenge Design While it is clear that dynamic enrollments can only rely on fixed challenges as no user involvement is possible, static enrollments can work with fixed, controlled or open challenges. We tested all three challenge types and found ambiguous results for fixed challenges. For example, many users had difficulties in figuring out how to draw complex sketches with two fingers simultaneously and could not reproduce them during enrollment or fallback authentication four weeks later. With respect to location-based questions, the fixed challenges were less problematic, but there were some users that selected questions that were related to less important episodes in their lives and thus forgot the answers to these questions

within six months. Overall, we observed that guided questions were the best option to encourage the definition of more secure questions through personal customization (e.g., by the inclusion of multiple pieces of information). Therefore, we recommend to use simple questions that keep the mental effort low, but that at the same time ask users (either implicitly or explicitly) for more than one piece of information to make it more difficult for potential adversaries to guess the correct answers (e.g., “*Where is my sister’s husband from?*”).

It is also advisable to divide the authentication task into small challenges that are consecutively shown to the user to reduce the mental effort, enabling users to focus on one thing at a time. This also has the advantage that the number of challenges required is more flexible and not everything is revealed at once [75]. For example, showing all items to be positioned for the icon-arrangements approach revealed to potential adversaries how many sub-tasks they have to solve and enabled them to infer relationships between the given items.

Response Design Similar to the design of challenges, responses can be fixed, controlled and open. For dynamic enrollments we mainly focused on fixed answers and advise to use them for cued-based recall as users are enrolled implicitly and may have problems to reproduce the exact answers otherwise. However, the use of fixed answers requires a certain amount of questions to be asked during fallback authentication to prevent random adversaries from succeeding by guessing. This number varies from approach to approach and depends on the number of answer options provided and the desired security level. For example, while a combination of seven questions about different smartphone activities with four answer options per question was sufficient to reach a comparable security level to PIN authentication, asking about installed apps required at least 40 questions to reach a similar security level.

We also tested guided/open answers for static enrollments, but the results were ambiguous: While they worked well for one approach (location-based questions), they did not for the other (2-finger sketches). The insights do not allow us to make a clear recommendation, but we want to encourage researchers not to shy away from the use of guided/open answers when novel techniques for answer input are explored. For example, we were able to obtain very promising results for our combination of location-based questions and map-based input.

7.3 Evaluating Fallback Authentication

The different design options for fallback authentication schemes and the potential of different types of autobiographical memories were evaluated with the help of different usability and security criteria (see Section 4.3). A summary of the key criteria is given next.

7.3.1 Usability Evaluation

Since we focused on knowledge-based schemes, memorability was one of the key usability factors to ensure the effectiveness of the corresponding fallback scheme. However, we also learned that the importance of user satisfaction should not be underestimated despite the infrequent use of fallback authentication.

Memorability Due to the natural process of forgetting it was not surprising to observe that the ability to recall authentication-related information decreased over time. But, it was interesting to learn about the reasons why this decrease happened. For example, with the help of repeated memorability tests at different points in time, we were able to find that errors for location-based questions were made due to ignorance of how the system works than due to forgetting the actual answers. From this observation, we conclude that repeated memorability tests can be beneficial for the evaluation of static enrollments to reveal usability issues on the one hand and to simulate a realistic fallback scenario on the other. There is no strict time schedule for memorability tests, but, based on our experience, three measurements seem advisable to reflect different fallback situations.

For example, a memorability test shortly after enrollment can help to quickly discover usability and memorability issues to improve the evaluated prototypes or to reject the approach completely. Another test after one to four weeks is useful to see if this information can be recalled after a certain time has passed and to analyze what kinds of errors users make (e.g., whether they make similar errors as for the first memorability test). It reflects situations in which users have just registered for a service/device and thus may get locked out due to the lack of training. While these short-term evaluations are a quick way to conclude whether a certain idea is worth pursuing, long-term evaluations (e.g., after six months or longer) are essential to simulate realistic fallback scenarios that allow better conclusions about the actual memorability of a system.

With respect to dynamic enrollments, long-term evaluations are not needed as they work with recent information. For example, we conducted only one memorability test after four weeks of logging for questions about smartphone activities. However, it would have also been interesting to see, if the availability of data for these activities would have changed, if we had conducted another test a few months after the first memorability test.

User Satisfaction We argued that user satisfaction plays only a minor role in the design of fallback schemes due to their infrequent use (see Chapter 4). While this is true to some extent, our results suggest that a certain level of user satisfaction must be met with respect to the learnability of new schemes. For example, although users understood the general idea of 2-finger sketches, they found it hard to learn how to actually draw them with two fingers simultaneously and thus developed undesired behaviors and were unsatisfied with the system as a whole. This means that fallback schemes that are hard to learn can prevent users from setting them up in the first place or can compromise their security due to undesired behaviors by the user.

7.3.2 Security Evaluation

For the security evaluations, we took into account different types of threats. We started with a baseline comparison with PIN authentication with respect to the theoretical security level of the corresponding fallback scheme. This was done to ensure that the fallback chain is not weakened when another component is added. In addition to this, the baseline allowed us to define important authentication parameters, such as the number of required answers or the number of allowed attempts.

We also took into account different types of human adversaries, but focused on attacks by close persons. Since these adversaries are more familiar with the victim than acquaintances or strangers, they are also more likely to know the answers to personal questions. Their threat potential is emphasized by our study results: Most adversaries that succeeded in their attacks were part of the actual memory and thus knew the answers (e.g., a joint vacation), were the answers themselves (e.g., communication partner) or were the creator of the answers (e.g., icon arrangements).

This shows that close adversaries are a good means to test the security of autobiographical memories. But since not all close adversaries have the same threat potential, it is beneficial for security evaluation to have a more refined view on this matter. Instead of categorizing adversaries by their closeness to the user only, one should also take into account the trust level that users have in them. This has the advantage to create a more realistic threat estimation. For example, trusted adversaries often have access to the user's device already and are not likely to attack the user. In turn, less trusted (but still close) adversaries are more likely to do so. Overall, we suggest the use of close and trusted adversaries to evaluate autobiographical memories from a worst-case perspective, but we also advise to complement this analysis with close and less trusted adversaries to get a more realistic security estimation.

7.3.3 Other Evaluation Criteria

In addition to the previously described criteria, we found additional factors in the course of our studies that have to be taken into account for the design of fallback schemes. This concerns, above all, the design of dynamic enrollments.

Privacy The main problem of dynamic enrollments is their use of potentially sensitive information (e.g., communication details or photos) and the disclosure of this information during fallback authentication without the explicit consent of the user. Therefore, users should be in control to decide whether a certain authentication scheme is activated or not (e.g., during setup), but, beyond that, the choice of data sources for the design of dynamic enrollments must be done carefully: the information used should be personal enough to distinguish between users, but at the same time not too private to be shared with others. We found that app-related information possesses these properties and thus is an interesting data source to be considered.

Data Availability Another problem of dynamic enrollments is related to data availability. We found in our studies that there are very different types of users: Some of them showed a high smartphone usage behavior, while others were the exact opposite. There were situations for the latter in which we could not create the authentication task due to the lack of data. Dynamic enrollments should take this into account and adapt the fallback scheme accordingly, for example, by using data sources for which enough data is available.

In addition to this, data availability can also become a problem when certain trends arise. For example, while music was mostly stored on one's device in the past, it is now often consumed through streaming services. Although APIs exist to retrieve information from these services, their use would not comply with the requirements for alternative fallback schemes (e.g., independency from Internet access). This means that dynamic fallback schemes must be updated to ensure that they work during lockout situations.

7.3.4 Order of Importance

Although the discussed criteria are all relevant for the design of fallback schemes, it is difficult to meet all of them at the same time. Therefore, compromises and prioritization need to be made. Since privacy is a deciding factor whether a fallback scheme is accepted or not, it should get the highest priority. This is followed by security and memorability factors that often have to be considered together to find a satisfying trade-off between the two. Last, fallback schemes should be easy to learn to achieve user satisfaction during the fallback experience.

7.3.5 Analyzing Fallback Authentication

For the analysis of our data, we used the accuracy metric to find the best trade-off between usability and security. It was a quick way to identify promising autobiographical memories on the one hand, and to reject autobiographical memories with less potential on the other. It further allowed us to define the best parameters (i.e., number of required questions and number of allowed attempts) for real-world deployments based on the best trade-off between memorability and security.

While accuracy is commonly used for the analysis of biometric authentication schemes (Section 4.3.3), it is less established for knowledge-based schemes. The latter mainly focus on success/failure rates of users and potential adversaries. These rates are indisputably important. However, since they are also part of the accuracy formula, little effort is required to complement these reports with the corresponding accuracy values. We advise to do this to enable better comparability.

7.4 Application Areas

While the focus of this thesis was fallback authentication on smartphones, it is an interesting question whether the results are also useful for other application areas. For example, tablets have become popular over the past few years. They have similar properties as smartphones, but, at the same time, are different due to the way they are used. Research from 2012 about tablets found that these devices are often used in home environments and shared with other family members [126]. Thus, in theory, the approaches in this work can also be applied to tablets, but, in practice, this cannot be done without limitations for several reasons:

First, once devices are shared, dynamic enrollments become difficult as usage data is no longer generated by one person only which, in turn, is likely to affect memorability as well as security (e.g., the revelation of data from another person). Second, the predominant use of tablets in home environments makes the need for alternative fallback schemes less important as there are quicker ways to regain access to the device (e.g., connecting to the computer).

However, there are, similar to smartphones, use cases in which alternative schemes are also beneficial on tablets (e.g., when lockouts happen on vacations). When dynamic fallback schemes are used in these situations, one has to pay particular attention that certain activities are not highlighted for other users. For example, although users of shared devices already have access to communication histories, they may not explicitly search for this information on the device. But when they encounter this information during fallback authentication, they may stumble on information they did not expect. Therefore, we advise not to use these kinds of information to prevent social tensions. Again, we see a great potential for app-related approaches, but whether users are able to tell whether certain apps are installed or not installed on shared devices needs yet to be tested.

7.5 Lessons Learned

In this chapter, we recapitulated the key results from this thesis and discussed their implications for the motivation, design, evaluation and analysis of fallback schemes on smartphones that exploit autobiographical memories. The insights from our exploration are helpful guidelines for usable security and privacy research into alternative authentication schemes. The key aspects that need to be taken into account are summarized in the following.

- **Authentication Chain.** Consider authentication as a chain of successively issued authentication schemes for which the following components are suggested: a primary scheme for regular authentication (e.g., PIN), a dynamic fallback scheme in case the primary scheme fails, and a static fallback scheme in case the dynamic scheme fails (e.g., [77, 143]; see Sections 3 and 4).

- **Use Case.** Know for which part of the chain you are designing to identify important requirements and to define convincing use cases. This sounds simple, but is often neglected, although it is crucial for the development of appropriate authentication schemes (see Section 4).
- **Fallback Scheme Type.** Use knowledge-based approaches for the design of fallback schemes as they are a better fit to accommodate the requirements for fallback authentication (i.e., independency, immediacy and engagement). Token-based schemes can also be used, but are better suitable as a last-resort option (see Sections 2 and 3).
- **Autobiographical Memories.** Design knowledge-based fallback schemes that focus on personal (but not private) episodes from the declarative memory. These memories must be actively created by the user to establish a strong spatial or temporal context that supports better memorization (e.g., [36, 74, 76, 77, 83]; see Sections 5 and 6).
- **Privacy Evaluation.** Use privacy as the key factor to decide whether a certain type of memory should be used for the design of an authentication scheme. If users are uncomfortable with sharing this memory with others, it should be neglected in favor of less private memories (e.g., [75]; see Section 6).
- **Task Design.** Reduce the mental effort needed for the retrieval of autobiographical memories by dividing the authentication task into smaller sub-tasks. Favor fixed answers for the design of dynamic enrollments, but do not hesitate to explore the potential of open answers for static enrollments to replace text-based input. In both cases, pay attention that a sufficient number of tasks are required and a sufficient number of attempts are allowed to achieve a certain level of usability and security (e.g., [77, 156]; see Section 6).
- **Usability Evaluation.** Evaluate user satisfaction at an early stage as it is an important factor for the acceptance of an authentication scheme (and thus allows a quick rejection of ideas when this requirement is not met). However, memorability is the key factor for the evaluation of knowledge-based schemes. For this, simulate a realistic fallback setting and test the ability to recall different types of memories at different points in time (e.g., [76, 100]; see Section 5).
- **Security Evaluation.** Evaluate the security of autobiographical memories with two adversary types: close adversaries that are trusted by the user and that have advanced knowledge about them (i.e., worst-case scenario) as well as adversaries that are close, but not necessarily trusted (e.g., [75–77, 103, 128]; see Sections 3, 4, 5 and 6).
- **Data Analysis.** Use the accuracy metric to compare different approaches to each other. Report the accuracy values, the success rates of attacks and the failure rates of authentication attempts. This allows to quickly identify promising approaches and to find the best trade-off between usability and security (e.g., [75, 121]; see Section 4).

8

Looking Back, Moving Forward

Remember the girl that locked her SIM card after a flight to Denmark? That girl was me, the author of this thesis. I was on my first travel as a research assistant and very excited about this trip, the conference and my talk. I got a little annoyed by the lockout incident as I was not able to share my experience with friends and family. At that time, SMS communication was still common and not possible without a working SIM card. I really felt lost as the alternative code to unlock it was at home, miles away, in one of my drawers. There is no need to worry, I quickly got over my annoyance and enjoyed the conference. But in retrospect, it feels like destiny that I spent almost four years of my research activities on mobile fallback authentication to support users in similar situations. I enjoyed every bit of this research and hope you did as well by reading this thesis. It was a long journey and thus deserves a final summary (Section 8.1) and a brief peak into the future (Sections 8.2 and 8.3).

8.1 Contribution Summary

Fallback authentication has always been the second choice. Not only in the way it is used, but also in the way it is treated by the research community: The design of new authentication schemes clearly sets its priorities on primary authentication. Only in cases where the schemes fail, fallback authentication is considered as another possibility. This has happened in the desktop environment (e.g., [190]) and also in the mobile environment (e.g., [37]). The goal of this thesis was to change this situation. In particular, we wanted to create a better understanding of fallback authentication to help it step out of its shadows. This is important to help researchers and practitioners to decide for which context of use they are designing before new authentication schemes are actually proposed, implemented and evaluated. The insights in this work contributed to this goal in several ways:

First, they motivated the need for mobile fallback authentication by understanding the problems of current solutions and the needs that users have during lockouts. Second, this understanding enabled us to create a fallback authentication framework that helps researchers to classify their ideas (i.e., to decide if they are designing for fallback authentication or primary authentication). Third, through the analysis of different types of autobiographical memories, we identified promising and less promising data sources that should or should not be used for the design of fallback schemes. Our example implementations provided a glimpse into the design space of fallback schemes and further demonstrated how these schemes should be evaluated in order to get helpful results in the context of fallback authentication.

8.1.1 Understanding Fallback Authentication

Within this thesis, we provided an overview of current solutions for fallback authentication that are used in desktop environments as well as mobile environments (see Chapters 1 and 2). It quickly became apparent that solutions for the latter were not well documented and little understanding on their usage was available. We are the first to provide insights on this matter by evaluating how often lockouts happen and how often they fail the user. It was not surprising that lockout occurrences are rare, but the problems that we identified through personal anecdotes by smartphone owners were interesting and motivated our work to design alternative fallback schemes that are independent, immediate and engaging (Chapter 3).

8.1.2 Definition of Fallback Authentication

During our research, we encountered different terminology that described the act to authenticate with an alternative scheme when the primary scheme fails (see Chapter 1). None of these descriptions went beyond a brief explanation. We therefore contributed to a better definition of fallback authentication (see Chapter 4) and modified it based on our study results. It can be summarized as follows:

Definition.

Fallback authentication is part of an authentication chain that consists of $1 \dots n$ authentication schemes, with $n \in \mathbb{N}$. Each authentication scheme in the chain involves a three-stage process: enrollment, authentication and replacement. An authentication scheme is called fallback scheme if it is preceded by another authentication scheme in the chain. Otherwise it is referred to as the primary scheme. Fallback schemes should further fulfill the following requirements to distinguish them from primary authentication schemes:

Effectiveness and Accuracy

Primary schemes must be effective and accurate in distinguishing between users and adversaries. This also applies for fallback schemes. Thus, fallback schemes must be at least as secure as the authentication scheme that precedes them to ensure a certain level of security.

Efficiency

Primary schemes must be efficient due to their frequent use, while fallback schemes can, but do not have to be due to their infrequent use. Thus, fallback schemes are less time-sensitive than the schemes that precede them, but more time-sensitive than the ones that follow them. This is done to reflect their frequency of use and to define their order of issue: Fallback schemes should be sorted by their required authentication times in ascending order.

Memorability

Primary schemes must be memorable to enable recall during authentication. This also applies for fallback schemes, but is aggravated due to the lack of training. Thus, fallback schemes must be memorable even without training. They must not rely on information that is learned by heart (e.g., through repetition), but should take advantage of recent memories or memories that are less likely to be forgotten over time (e.g., episodic memories).

Learnability and Satisfaction

Primary schemes must be learnable and satisfying to prevent dropouts and the deactivation of security measures, while fallback schemes must be learnable and satisfying to be usable during lockout situations: Users should immediately know how to authenticate when they encounter the scheme for the first time. Thus, fallback schemes must be easy to learn due to their infrequent use to satisfy users.

Motivation

Primary schemes exist to motivate users to protect their data, while fallback schemes exist to mitigate potential problems of primary schemes. Thus, fallback schemes must not lead to the loss of data to maintain the goals of primary schemes that they substitute. The enrollment effort for fallback schemes must be kept to a minimum to motivate users to set them up.

8.1.3 Autobiographical Memories

Based on the insights from the preceding contributions, we exploited autobiographical memories for the design of fallback schemes as they fulfilled well the requirements for mobile fallback authentication: They are independent (e.g., from other persons) as they solely rely on the user, they are immediate as they are stored in the user's brain (which is hopefully always with the user), and they are engaging as some of these memories (e.g., episodic memories) are remembered vividly to relive the past [175].

Our contribution is that we extended the design space of autobiographical memories by taking into account memory types that go beyond personal facts, such as episodic memories or procedural skills (see Chapter 2). Their evaluation in five example projects enabled us to identify their potential and challenges in terms of memorability and security in the context of fallback authentication (see Chapters 5 and 6). Episodic memories with a strong location-dependent context showed the best results in terms of memorability and security: While almost all legitimate users were able to authenticate, none of the close adversaries or strangers were able to attack these memories successfully. App-related activities also showed good usability and security properties when a certain amount of these memory types was available. But in addition to this, we also identified less suitable memory types, such as memories about photos, music or drawings. These memory types failed for different reasons that were related to memorability, privacy or user satisfaction. We showed that autobiographical memories that were actively created by the user and that had a strong spatio-temporal context with a personal, but not private notion, were the most promising memory types.

8.1.4 Design of Fallback Schemes

The evaluation of different autobiographical memories was done through the example implementation of fallback schemes. This means that the memorability and security of the memories are also affected by the way they were implemented. For example, our location-based questions worked very well due to the required input on a map, while similar questions from related research that used text-based answers had much worse recall rates due to repeatability issues (e.g., [139]).

Therefore, this thesis also contributes to the design space of fallback schemes. We summarized different authentication-related classifications from related work (e.g., [101, 187]) and derived the design space for fallback authentication based on them. Hereby, our focus was on the design and implementation of static and dynamic enrollments in combination with different types of autobiographical memories (see Chapters 5 and 6). This enabled us to highlight and discuss the advantages and disadvantages of different design options based on which we were able to make suggestions about how they can be applied in the fallback chain. For example, instead of considering static and dynamic enrollments as competing schemes, it is advisable to include them both into the fallback chain to mitigate each other's shortcomings and to act as each other's fallback solution.

8.1.5 Evaluation of Fallback Schemes

A major part of this thesis was the evaluation of autobiographical memories and the example fallback schemes using a consistent research methodology throughout all projects. Our main intention was to demonstrate how memorability and security can be evaluated in the context of fallback authentication with respect to autobiographical memories. We further strived for the identification of other important criteria that must be considered for evaluation.

The example fallback schemes with static enrollments showed that it is important to use repeated memorability tests to have a realistic simulation for fallback authentication. While early tests allowed the quick rejection of ideas (e.g., 2-finger sketches), long-term evaluations provided better insights into the stability/change of recall rates over time (e.g., location-based questions). The use of multiple tests was of no concern for dynamic enrollments, but their evaluation showed that the suitability of different types of autobiographical memories was not only a trade-off between usability and security, but also a trade-off with privacy.

Within our five projects, we used the accuracy metric to find the best balance between memorability and security. This approach was not new, but still uncommon for knowledge-based schemes. We provided reasons for and against the use of this metric and demonstrated its use throughout all of our projects.

8.2 Limitations and Future Work

Although there was some previous knowledge from related work that we could build upon, fallback authentication, in particular on mobile devices, was still an unexplored domain that left us with many options that had to be weighted up against each other in order to make the right decisions with respect to the selection of autobiographical memories, the design of the fallback schemes as well as their evaluation. Our choices were carefully made by taking into account previous work; where no preceding work was available, we followed an iterative design approach to reject bad ideas early and to pursue more promising directions. Nonetheless, this does not save our work from certain limitations that we would like to use to light the way for future research in the domain of fallback authentication.

8.2.1 In-Depth Insights on Mobile Fallback Authentication

While we believe that our online survey reflects well how rarely fallback authentication happens, we also believe that the individual reports did not unveil all potential problems of lockout experiences due to the small sample size. Since our work was the first to address the topic of mobile fallback authentication, we were striving for a broad overview on this matter to set the basis for in-depth evaluations in the future. Still, the problems that were unveiled, provided a good motivation for the alternative schemes proposed in this thesis.

8.2.2 Different User Types

With respect to these schemes, one might argue that they are only helpful for a small target group. It is true that our implementations were targeted at a very specific audience, but the availability of alternatives can also be beneficial for those who did not report major problems with current solutions, but who expressed annoyance with them. Nonetheless, there are limitations to the generalizability of the results on different levels:

For example, the study results in chapter 6 revealed that smartphone owners have very different usage patterns. In particular, owners that seldom use their smartphone may not generate enough data for implicit enrollments. It would be interesting to analyze this in more detail by testing how well the proposed approaches work between different groups that exhibit different smartphone usage patterns.

In addition to this, our study samples were very young so that we cannot rule out a possible influence of age on the recall rates. However, this is, to some extent, mitigated by our choice to take advantage of very recent memories or by allowing users to define their own questions. For example, younger and older persons may have different associations to certain guidelines for location-based questions (see Chapter 5), but the possibility of customization enables different types of users to define relevant questions for them to ensure applicability.

8.2.3 Development Platform

The research presented in this thesis mainly focused on fallback authentication on smartphones. We had discussed potential application areas outside this context (see Chapter 7), but there is also the question how well the results can be transferred to other mobile platforms. Although smartphones with different operating systems have their distinct features, most of them allow users to do similar activities, such as making calls, using apps, install apps or personalize their devices. This means that the presented ideas in this work can also work for non-Android devices, but have to be adapted to manufacturer-specific features. For example, iPhone users do not have the possibility to place Widgets on their home screens, while Windows Phone 7 users have so-called Live Tiles for home screen organization.

8.2.4 Beyond Episodic Memories

Within this thesis, we showed that autobiographical memories are more than just personal facts and thus had a closer look at the potential of episodic memories that referred to the very past (Chapter 5) or to recent activities (Chapter 6). However, during the evaluation of location-based questions, we made an interesting observation that questions for fallback authentication can also refer to the future. This is an interesting direction for research. To some extent, this has been done for plans in the near future when, for example, dynamic questions are generated from calendar entries (e.g., [6]). But the study of plans for the distant future

has yet to be done. On the one hand, these kinds of plans have the advantage that they are potentially relevant and important for the user, but on the other hand, they have the disadvantage that plans are also prone to change so that questions about them may be outdated when fallback authentication takes place. But besides future plans, autobiographical memory further contains other types of memories about the past that go beyond personal facts as well as episodic memories. This includes, for example, the non-declarative memory that is also responsible for conditioning and priming. For the latter, Denning et al. have shown that priming has a potential for the design of fallback schemes, so that this direction for research is also interesting [46].

8.2.5 Real-World Deployment

In the context of this work, we have analyzed the advantages and disadvantages of different types of episodic memories. Location-based questions showed very promising results in terms of memorability and security so that we trust in their potential to replace traditional security questions when some changes are made (i.e., the verification of the answer during enrollment). For this, we hope that location-based questions will be deployed by some service providers to test their feasibility in the real world.

With respect to dynamic enrollments, we were able to show that app-related questions have the best potential for the design of fallback schemes. However, the number of false positives and false negatives prevents us from recommending the fallback schemes, as is, for a real-world deployment for the time being. Instead of evaluating the different data sources individually, we suggest to combine them to build and evaluate them with real-world parameters. This means that users are not asked about one data source for fallback authentication, but about all of them: Instead of answering many questions within a question category, users answer few questions from many categories that, for example, do not only include questions about their app usage, but also about the apps that they install or have installed on their smartphones for fallback authentication.

8.3 Closing Remarks

Having been the second choice, neglected, underestimated or even ignored by the research community, fallback authentication had a cruel fate. We wanted to raise awareness for its existence and, hopefully, we succeeded. It was a long journey with setbacks, but mostly with interesting insights that guided us along the way and that kept our strive. For us, this journey ends here, but it will hopefully be continued by other members of the research community. We believe that authentication in general, will always be an important topic due to the increasing amount of explicit and implicit information that we store and create with our technical devices and that, in turn, require some kind of protection to prevent unauthorized access. And as long as authentication exists, fallback authentication will be a part of it. It

is up to the researchers and practitioners of the usable security and privacy community to decide how this co-existence will turn out for fallback authentication. Remember: Always know for which authentication context you are designing for and set your design focus on fallback authentication where suitable.

A

Appendices

A.1 Overview of Fallback Schemes in 2014

	URL	ER	EA	SQ	AA	TC	P	CS	O	Comments
1	Google.com	✓	✓		✓		✓			
2	Facebook.com	✓	✓	✓		✓			✓	personal data (i.e. ID)
3	Yahoo.com	✓		✓		✓	✓			
4	Amazon.com	✓						✓		
5	Wikipedia.org	✓								
6	Twitter.com	✓					✓	✓		
7	Live.com	✓		✓			✓			
8	Linkedin.com	✓								
9	Ebay.com	✓						✓		
10	Adcash.com	✓								
11	Instagram.com	✓							✓	facebook account
12	Reddit.com	✓								
13	Ask.com	✓					✓			
14	Tumblr.com	✓					✓			
15	Pinterest.com	✓								
16	Wordpress.com	✓						✓		command line tools
17	Paypal.com	✓		✓						
18	Imgur.com	✓								
19	Aliexpress.com	✓								
20	Xvideos.com	✓								
21	Alibaba.com	✓								
22	Apple.com	✓		✓					✓	trusted device
23	Imdb.com	✓			✓					
24	go.com	✓								
25	onclickads.net	✓								
26	Netflix.com	✓			✓					
27	Xhamster.com	✓								
28	Stackoverflow.com	✓								
29	kickass.so	✓								
30	Craigslist.org	✓								
31	Googleadservice.com		✓		✓		✓			
32	Adobe.com	✓						✓		
33	Pornhub.com	✓								
34	Espn.go.com	✓								
35	bbc.co.uk	✓								
36	Odnoklassniki.ru	✓						✓		
37	cnn.com	✓								

38	dailymotion.com	✓	
39	flipkart.com	✓	
40	walmart.com	✓	
41	dropbox.com	✓	
42	huffingtonpost.com	✓	
43	wiki.com	✓	

Table A.1: Overview of fallback schemes of 43 of the 100 top Alexa websites in December 2014. ER = Email-based reset based; EA=Email-based reset with alternate email; SQ = security questions; AA = account activity; TC = trusted contacts; P = phone-based reset; CS = customer service; O = other.

Bibliography

- [1] 3rd Generation Partnership Project 2. cdma2000 application on UICC for spread spectrum systems. *Standard C.S0065-0 (version 2.0)*, 2008.
- [2] A. Acquisti and R. Gross. Imagined communities: Awareness, information sharing, and privacy on the Facebook. In *Proceedings of the 6th International Conference on Privacy Enhancing Technologies (PET'06)*. Springer, 2006, pages 36–58.
- [3] A. Adams and M. A. Sasse. Users are not the enemy. *Communications of the ACM*, Vol. 42(12), ACM, 1999, pages 40–46.
- [4] S. Al-Zubi, A. Brömmel, and K. Tönnies. Using an active shape structural model for biometric sketch recognition. *Pattern Recognition*, Vol. 2781, Springer, 2003, pages 187–195.
- [5] R. C. Atkinson and R. M. Shiffrin. Human memory: A proposed system and its control processes. In R. C. Atkinson, R. M. Shiffrin, and K. W. Spence (editors). *The psychology of learning and motivation: Advances in research and theory*. Academic Press, 1968, pages 89 – 105.
- [6] A. Babic, H. Xiong, D. Yao, and L. Iftode. Building robust authentication systems with activity-based personal questions. In *Proceedings of the 2nd ACM workshop on Assurable and usable security configuration*. ACM, 2009, pages 19–24.
- [7] A. D. Baddeley. What is memory? In A. D. Baddeley (editor). *Essentials of Human Memory*. Psychology Press, 1999, pages 1–18.
- [8] A. D. Baddeley. The psychology of memory. In A. D. Baddeley, M. Kopelman, and B. A. Wilson (editors). *The Essential Handbook of Memory Disorders for Clinicians*. John Wiley & Sons, 2004, pages 1–14.
- [9] J. Bentley and C. Mallows. How much assurance does a pin provide? *Human Interactive Proofs*, Vol. 3517, Springer, 2005, pages 111–126.
- [10] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. *Workshop on Knowledge Discovery in Databases*, Vol. 10(16), AAAI Press, 1994, pages 359–370.
- [11] D. Berntsen. The unbidden past involuntary autobiographical memories as a basic mode of remembering. *Current Directions in Psychological Science*, Vol. 19(3), SAGE Publications, 2010, pages 138–142.
- [12] R. Biddle, S. Chiasson, and P. Van Oorschot. Graphical passwords: Learning from the first twelve years. *ACM Computing Surveys*, Vol. 44(4), ACM, 2012, pages 19:1–19:41.

-
- [13] M. Böhmer, B. Hecht, J. Schöning, A. Krüger, and G. Bauer. Falling asleep with angry birds, facebook and kindle: A large scale study on mobile application usage. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI'11)*. ACM, 2011, pages 47–56.
 - [14] M. Böhmer and A. Krüger. A study on icon arrangement by smartphone users. In *Proceedings of the 30th International Conference on Human Factors in Computing Systems (CHI'13)*. ACM, 2013, pages 2137–2146.
 - [15] J. Bonneau. The science of guessing: Analyzing an anonymized corpus of 70 million passwords. In *Proceedings of the 33rd IEEE Symposium on Security and Privacy (SP'12)*. IEEE, 2012, pages 538 – 552.
 - [16] J. Bonneau, M. Just, and G. Matthews. What’s in a name? Evaluating statistical attacks on personal knowledge questions. *Financial Cryptography and Data Security*, Vol. 6052, Springer, 2010, pages 98–113.
 - [17] J. Bonneau and S. Preibusch. The password thicket: Technical and market failures in human authentication on the web. *Workshop on the Economics of Info Security*, 2010.
 - [18] J. Bonneau, S. Preibusch, and R. Anderson. A birthday present every eleven wallets? The security of customer-chosen banking pins. *Financial Cryptography and Data Security*, Vol. 7397, Springer, 2012, pages 25–40.
 - [19] J. Brainard, A. Juels, R. L. Rivest, M. Szydlo, and M. Yung. Fourth-factor authentication: Somebody you know. In *Proceedings of the Conference on Computer and Communications Security (CCS '06)*. ACM, 2006, pages 168–178.
 - [20] S. Brostoff and M. A. Sasse. Are passfaces more usable than passwords? A field trial investigation. *People and Computers XIV - Usability or Else!*, Springer, 2000, pages 405–424.
 - [21] R. Brunelli and D. Falavigna. Person identification using multiple cues. *Pattern Analysis and Machine Intelligence*, Vol. 17(10), IEEE, 1995, pages 955–966.
 - [22] A. Buchoux and N. L. Clarke. Deployment of keystroke analysis on a smartphone. In *Proceedings of the 6th Australian Information Security Management Conference. secAU*, 2008, pages 48–55.
 - [23] J. Bunnell, J. Podd, R. Henderson, R. Napier, and J. Kennedy-Moffat. Cognitive, associative and conventional passwords: Recall and guessing rates. *Computers & Security*, Vol. 16(7), Elsevier, 1997, pages 629 – 641.
 - [24] C. Castelluccia, M. Dürmuth, and D. Perito. Adaptive password-strength meters from markov models. In *Proceedings of the 19th Annual Network & Distributed System Security Symposium (NDSS'12)*. The Internet Society, 2012.

- [25] J.-C. Chebat and P. Filiatrault. The impact of waiting in line on consumers. *International Journal of Bank Marketing*, Vol. 11(2), MCB UP Ltd, 1993, pages 35–40.
- [26] C.-H. Chen and C. T. Chu. Fusion of face and iris features for multimodal biometrics. *Advances in Biometrics*, Vol. 3832, Springer, 2005, pages 571–580.
- [27] 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 Conference on Computer and Communications Security (CCS '09)*. ACM, 2009, pages 500–511.
- [28] S. Chiasson, P. C. van Oorschot, and R. Biddle. A usability study and critique of two password managers. In *Proceedings of the 15th Conference on USENIX Security Symposium (USENIX-SS'06)*. USENIX Association, 2006.
- [29] S. Chiasson, P. C. van Oorschot, and R. Biddle. Graphical password authentication using cued click points. *Computer Security - ESORICS 2007*, Vol. 4734, Springer, 2007, pages 359–374.
- [30] G. Chittaranjan, J. Blom, and D. Gatica-Perez. Who's who with big-five: Analyzing and classifying personality traits with smartphones. In *Proceedings of the 15th International Symposium on Wearable Computers (ISWC'11)*. IEEE, 2011, pages 29–36.
- [31] P. Conrad. It's boring: Notes on the meanings of boredom in everyday life. *Qualitative Sociology*, Vol. 20(4), Kluwer Academic Publishers-Plenum Publishers, 1997, pages 465–475.
- [32] M. Conway and C. Pleydell-Pearce. The construction of autobiographical memories in the self memory system. *Psychological Review*, Vol. 107(2), APA, 2000, pages 261–288.
- [33] M. A. Conway. Sensory-perceptual episodic memory and its context: Autobiographical memory. In A. Baddeley, J. Aggleton, and M. Conway (editors). *Episodic Memory: New Directions in Research*. Oup Oxford, 2002, pages 1375–1384.
- [34] F. J. Corbató, J. H. Saltzer, and C. T. Clingen. Multics: The first seven years. In *Proceedings of the Spring Joint Computer Conference (AFIPS '72)*. ACM, 1972, pages 571–583.
- [35] J. E. Cutting and L. T. Kozlowski. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, Vol. 9(5), Springer, 1977, pages 353–356.
- [36] S. K. Dandapat, S. Pradhan, B. Mitra, R. Roy Choudhury, and N. Ganguly. Activ-pass: Your daily activity is your password. In *Proceedings of the 33rd International Conference on Human Factors in Computing Systems (CHI '15)*. ACM, 2015, pages 2325–2334.

-
- [37] A. Das, J. Bonneau, M. Caesar, N. Borisov, and X. Wang. The tangled web of password reuse. In *Proceedings of the 21st Annual Network & Distributed System Security Symposium (NDSS'14)*. Internet Society, 2014.
 - [38] 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*, Vol. 63(1), Elsevier, 2005, pages 128–152.
 - [39] A. De Luca, B. Frauendienst, S. Boring, and H. Hussmann. My phone is my keypad: Privacy-enhanced PIN-entry on public terminals. In *Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction (OZCHI '09)*. ACM, 2009, pages 401–404.
 - [40] 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 International Conference on Human Factors in Computing Systems (CHI '12)*. ACM, 2012, pages 987–996.
 - [41] 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 International Conference on Human Factors in Computing Systems (CHI '14)*. ACM, 2014, pages 1411–1414.
 - [42] A. De Luca, E. von Zezschwitz, and H. Hussmann. Vibrapass: Secure authentication based on shared lies. In *Proceedings of the International Conference on Human Factors in Computing Systems (CHI'09)*. ACM, 2009, pages 913–916.
 - [43] A. De Luca, R. Weiss, and H. Hussmann. PassShape: Stroke based shape passwords. In *Proceedings of the Australasian Conference on Computer-Human Interaction (OZCHI '07)*. ACM, 2007, pages 239–240.
 - [44] J. De Vriendt. Mobile network evolution: A revolution on the move. *Communications Magazine*, IEEE, 2002, pages 104–111.
 - [45] M. Demmler. Fallback authentication based on installed apps. *Bachelor Thesis at the University of Munich (LMU)*, 2014.
 - [46] T. Denning, K. Bowers, M. van Dijk, and A. Juels. Exploring implicit memory for painless password recovery. In *Proceedings of the Internatinal Conference on Human Factors in Computing Systems (CHI '11)*. ACM, 2011, pages 2615–2618.
 - [47] R. Dhamija and A. Perrig. Deja vu: A user study using images for authentication. In *Proceedings of the USENIX Security Symposium (SSYM'00)*. USENIX Association, 2000, page 4.

- [48] J. M. DiMicco and D. R. Millen. Identity management: Multiple presentations of self in facebook. In *Proceedings of the International Conference on Supporting Group Work (GROUP '07)*. ACM, 2007, pages 383–386.
- [49] A. E. Dirik, N. Memon, and J.-C. Birget. Modeling user choice in the passpoints graphical password scheme. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS'07)*. ACM, 2007, pages 20–28.
- [50] P. Dunphy and J. Yan. Do background images improve “Draw a Secret” graphical passwords? In *Proceedings of the Conference on Computer and Communications Security (CCS'07)*. ACM, 2007, pages 36–47.
- [51] H. Ebbinghaus. Memory: A contribution to experimental psychology. *Annals of Neurosciences*, Vol. 20(4), 2013, pages 155–156.
- [52] S. Egelman, S. Jain, R. S. Portnoff, K. Liao, S. Consolvo, and D. Wagner. Are you ready to lock? In *Proceedings of the Conference on Computer and Communications Security (CCS'2014)*. ACM, 2014, pages 750–761.
- [53] 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 International Conference on Human Factors in Computing Systems (CHI'13)*. ACM, 2013, pages 2379–2388.
- [54] B. D. Ehret. Learning where to look: Location learning in graphical user interfaces. In *Proceedings of the International Conference on Human Factors in Computing Systems (CHI '02)*. ACM, 2002, pages 211–218.
- [55] C. Ellison, C. Hall, R. Milbert, and B. Schneier. Protecting secret keys with personal entropy. *Future Generation Computer Systems*, Vol. 16(4), Elsevier, 2000, pages 311–318.
- [56] D. Florêncio and C. Herley. A large-scale study of web password habits. In *Proceedings of the International Conference on World Wide Web (WWW '07)*. ACM, 2007, pages 657–666.
- [57] D. Florêncio and C. Herley. Where do security policies come from? In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS'10)*. ACM, 2010, pages 10:1–10:14.
- [58] D. Florêncio, C. Herley, and B. Coskun. Do strong web passwords accomplish anything? In *Proceedings of the USENIX Workshop on Hot Topics in Security (HOT-SEC'07)*. USENIX Association, 2007, pages 10:1–10:6.
- [59] N. Frykholm and A. Juels. Error-tolerant password recovery. In *Proceedings of the Conference on Computer and Communications Security (CCS '01)*. ACM, 2001, pages 1–9.

-
- [60] S. Furnell. An assessment of website password practices. *Computers & Security*, Vol. 26(7-8), Elsevier, 2007, pages 445 – 451.
 - [61] D. Gafurov, K. Helkala, and T. Søndrol. Biometric gait authentication using accelerometer sensor. *Journal of Computers*, Vol. 1(7), Academy Publisher, 2006, pages 51–59.
 - [62] J.-L. Gao. Fallback authentication based on icon arrangement. *Bachelor Thesis at the University of Munich (LMU)*, 2013.
 - [63] S. L. Garfinkel. Email-based identification and authentication: An alternative to PKI? *Security & Privacy*, Vol. 1(6), Computer Society, 2003, pages 20–26.
 - [64] S. Gaw and E. W. Felten. Password management strategies for online accounts. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS '06)*. ACM, 2006, pages 44–55.
 - [65] T. Giorgino. Computing and visualizing dynamic time warping alignments in R: the DTW package. *Journal of Statistical Software*, Vol. 31(7), Foundation for Open Access Statistics, 2009, pages 1–24.
 - [66] J. Goldberg, J. Hagman, and V. Sazawal. Doodling our way to better authentication. In *Extended Abstracts on Human Factors in Computing Systems (CHI'02)*. ACM, 2002, pages 868–869.
 - [67] N. S. Govindarajulu and S. Madhvanath. Password management using doodles. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI'07)*. ACM, 2007, pages 236–239.
 - [68] V. Griffith and M. Jakobsson. Messin’ with texas deriving mother’s maiden names using public records. Vol. 3531, 2005, pages 91–103.
 - [69] J. Grossklags and A. Acquisti. When 25 cents is too much: An experiment on willingness-to-sell and willingness-to-protect personal information. In *Workshop on the Economics of Information Security*, 2007.
 - [70] P. Gupta, S. Gottipati, J. Jiang, and D. Gao. Your love is public now: Questioning the use of personal information in authentication. In *Proceedings of the Symposium on Information, Computer and Communications Security (ASIA CCS '13)*. USENIX Association, 2013, pages 49–60.
 - [71] S. Gustafson, D. Bierwirth, and P. Baudisch. Imaginary interfaces: Spatial interaction with empty hands and without visual feedback. In *Proceedings of the Symposium on User Interface Software and Technology (UIST'10)*. ACM, 2010, pages 3–12.
 - [72] W. J. Haga and M. Zviran. Question-and-answer passwords: An empirical evaluation. *Information Systems*, Vol. 16(3), Elsevier, 1991, pages 335–343.

- [73] A. Hang, A. De Luca, J. Hartmann, and H. Hussmann. Oh app, where art thou? On app launching habits of smartphone users. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '13)*. ACM, 2013, pages 392–395.
- [74] A. Hang, A. De Luca, and H. Hussmann. Using icon arrangement for fallback authentication on smartphones. In *Extended Abstracts on Human Factors in Computing Systems (CHI '14)*. ACM, 2014, pages 2467–2472.
- [75] A. Hang, A. De Luca, and H. Hussmann. I know what you did last week! Do you? Dynamic security questions for fallback authentication on smartphones. In *Proceedings of the International Conference on Human Factors in Computing Systems (CHI '15)*. ACM, 2015, pages 1383–1392.
- [76] A. Hang, A. De Luca, M. Richter, M. Smith, and H. Hussmann. Where have you been? Using location-based security questions for fallback authentication. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS'15)*. ACM, 2015, pages 169–183.
- [77] 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 International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '15)*. ACM, 2015, pages 295–305.
- [78] A. Hang, E. Von Zezschwitz, A. De Luca, and H. Hussmann. Too much information! User attitudes towards smartphone sharing. In *Proceedings of the Nordic Conference on Human-Computer Interaction (NordiCHI'12)*. ACM, 2012, pages 284–287.
- [79] I.-H. Hann, K.-L. Hui, S. T. Lee, and I. P. L. Png. Online information privacy: Measuring the cost-benefit trade-off. In *Proceedings of the International Conference on Information Systems (ICIS'02)*. Association for Information Systems, 2002, pages 1–10.
- [80] 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 *Proceedings of the Symposium on Usable Privacy and Security (SOUPS'14)*. USENIX Association, 2014, pages 213–230.
- [81] P. Hauptmann. Context-based security questions for backup authentication on smartphones. *Bachelor Thesis at the University of Munich (LMU)*, 2013.
- [82] E. Hayashi, R. Dhamija, N. Christin, and A. Perrig. Use your illusion: Secure authentication usable anywhere. In *Proceedings of the Symposium on Usable Privacy and security (SOUPS'08)*. USENIX Association, 2008, pages 35–45.

-
- [83] E. Hayashi and J. Hong. A diary study of password usage in daily life. In *Proceedings of the International Conference on Human Factors in Computing Systems (CHI'11)*. ACM, 2011, pages 2627–2630.
 - [84] C. Herley. Passwords: If we're so smart, why are we still using them? *Financial Cryptography and Data Security*, Vol. 5628, Springer, 2009, pages 230–237.
 - [85] C. Herley. So long, and no thanks for the externalities: The rational rejection of security advice by users. In *Proceedings of the Workshop on New Security Paradigms Workshop (NSPW '09)*. ACM, 2009, pages 133–144.
 - [86] C. Herley and P. C. van Oorschot. A research agenda acknowledging the persistence of passwords. *Security & Privacy*, Vol. 10(1), IEEE, 2012, pages 28–36.
 - [87] A. Herlitz, L.-G. Nilsson, and L. Bäckman. Gender differences in episodic memory. *Memory & Cognition*, Vol. 25(6), Springer, 1997, pages 801–811.
 - [88] S. Hinduja and J. W. Patchin. Personal information of adolescents on the internet: A quantitative content analysis of myspace. *Journal of Adolescence*, Vol. 31(1), Elsevier, 2008, pages 125 – 146.
 - [89] L. Hirschbeck. Sketch-based authentication on smartphones. *Bachelor Thesis at the University of Munich (LMU)*, 2013.
 - [90] S. Hoeffler and D. Ariely. Constructing stable preferences: A look into dimensions of experience and their impact on preference stability. *Journal of Consumer Psychology*, Vol. 8(2), Elsevier, 1999, pages 113–139.
 - [91] L. Hong and A. Jain. Integrating faces and fingerprints for personal identification. *Pattern Analysis and Machine Intelligence*, Vol. 20(12), IEEE, 1998, pages 1295–1307.
 - [92] B. Huberman, E. Adar, and L. Fine. Valuating privacy. *Security & Privacy*, Vol. 3(5), IEEE, 2005, pages 22–25.
 - [93] P. G. Inglesant and M. A. Sasse. The true cost of unusable password policies: Password use in the wild. In *Proceedings of the International Conference on Human Factors in Computing Systems (CHI '10)*. ACM, 2010, pages 383–392.
 - [94] I. Irakleous, S. Furnell, P. Dowland, and M. Papadaki. An experimental comparison of secret-based user authentication technologies. *Information Management & Computer Security*, Vol. 10(3), MCB UP, 2002, pages 100–108.
 - [95] M. Jakobsson, E. Shi, P. Golle, and R. Chow. Implicit authentication for mobile devices. In *Proceedings of the Conference on Hot Topics in Security (HOTSEC'09)*. USENIX Association, 2009, pages 9–15.

- [96] M. Jakobsson and H. Siadati. Improved visual preference authentication. In *Proceedings of the Workshop on Socio-Technical Aspects in Security and Trust (STAST'12)*. IEEE, 2012, pages 27 – 34.
- [97] M. Jakobsson, E. Stolterman, S. Wetzel, and L. Yang. Love and authentication. In *Proceedings of the International Conference on Human Factors in Computing Systems (CHI'08)*. ACM, 2008, pages 197–200.
- [98] I. Jermyn, A. J. Mayer, F. Monroe, M. K. Reiter, A. D. Rubin, et al. The design and analysis of graphical passwords. In *Proceedings of the USENIX Security Symposium (USENIX Security'99)*. USENIX Association, 1999, pages 1–15.
- [99] A. N. Joinson. Looking at, looking up or keeping up with people? motives and use of facebook. In *Proceedings of the International Conference on Human Factors in Computing Systems (CHI '08)*. ACM, 2008, pages 1027–1036.
- [100] M. Just. Designing and evaluating challenge-question systems. *Security & Privacy*, Vol. 2(5), IEEE, 2004, pages 32–39.
- [101] M. Just. Designing authentication systems with challenge questions. In L. F. Cranor and S. Garfinkel (editors). *Security and Usability: Designing Secure Systems That People Can Use*. O'Reilly, 2005, pages 143–155.
- [102] M. Just and D. Aspinall. Challenging challenge questions. In *Proceedings of the International Conference on Trust & Trustworthy Computing (Trust'09)*. Springer, 2009, pages 6–8.
- [103] M. Just and D. Aspinall. Personal choice and challenge questions: A security and usability assessment. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS'09)*. USENIX Association, 2009, Article No. 8.
- [104] R. Kainda, I. Flechais, and A. Roscoe. Security and usability: Analysis and evaluation. In *Proceedings of the International Conference on Availability, Reliability, and Security (ARES'10)*. IEEE, 2010, pages 275–282.
- [105] A. K. Karlson, A. B. Brush, and S. Schechter. Can i borrow your phone?: Understanding concerns when sharing mobile phones. In *Proceedings of the International Conference on Human Factors in Computing Systems (CHI '09)*. ACM, 2009, pages 1647–1650.
- [106] A. Karole, N. Saxena, and N. Christin. A comparative usability evaluation of traditional password managers. In *Proceedings of the International Conference of Information Security and Cryptology (ICISC'10)*. Springer, 2011, pages 233–251.
- [107] P. G. Kelley. Conducting usable privacy and security studies with Amazon's mechanical turk. *Panel at the Symposium on Usable Privacy and Security*, 2010.

-
- [108] H. Kim and J. H. Huh. PIN selection policies: Are they really effective? *Computers & Security*, Vol. 31(4), Elsevier, 2012, pages 484–496.
 - [109] H. Kim, J. Tang, and R. Anderson. Social authentication: Harder than it looks. *Financial Cryptography and Data Security*, Springer, 2012, pages 1–15.
 - [110] E. A. Kolek and D. Saunders. Online Disclosure: An Empirical Examination of Undergraduate Facebook Profiles. *Journal of Student Affairs Research and Practice*, 45(1), De Gruyter, 2008.
 - [111] 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 International Conference on Human Factors in Computing Systems (CHI'11)*. ACM, 2011, pages 2595–2604.
 - [112] C. Lampe, N. Ellison, and C. Steinfield. A Face(book) in the crowd: Social searching vs. social browsing. In *Proceedings of the International Conference on Computer Supported Cooperative Work (CSCW '06)*. ACM, 2006, pages 167–170.
 - [113] J. Leggett, G. Williams, M. Usnick, and M. Longnecker. Dynamic identity verification via keystroke characteristics. *International Journal of Man-Machine Studies*, Vol. 35(6), Elsevier, 1991, pages 859–870.
 - [114] D. Malone and K. Maher. Investigating the distribution of password choices. In *Proceedings of the International Conference on World Wide Web (WWW'12)*. ACM, 2012, pages 301–310.
 - [115] J. Mandler, D. Seegmiller, and J. Day. On the coding of spatial information. *Memory & Cognition*, Vol. 5(1), Springer, 1977, pages 10–16.
 - [116] M. Mannan, D. Barrera, C. D. Brown, D. Lie, and P. C. Van Oorschot. Mercury: Recovering forgotten passwords using personal devices. In *Financial Cryptography and Data Security*. Springer, 2012, pages 315–330.
 - [117] M. Mannan and P. C. van Oorschot. Security and usability: The gap in real-world online banking. In *Proceedings of the Workshop on New Security Paradigms (NSPW'07)*. ACM, 2008, pages 1–14.
 - [118] S. Marcel, C. Cool, C. Atanasoaei, F. Tarsetti, J. Pesán, P. Matejka, J. Cernocky, M. Helistekangas, and M. Turtinen. Mobio: Mobile biometric face and speaker authentication. *EPFL-REPORT No. 150602*, 2010.
 - [119] E. J. Marsh and H. L. Roediger. Episodic and autobiographical memory. In H. L. Roediger and E. J. Marsh (editors). *Handbook of Psychology*. Wiley, 2012.

- [120] 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 Conference on Computer and Communications Security (CCS'13)*. ACM, 2013, pages 173–186.
- [121] C. E. Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, Vol. 8(4), 1978, pages 283–298.
- [122] A. Moallem. Did you forget your password? *Design, User Experience and Usability. Theory, Methods, Tools and Practice*, Vol. 6770, Springer, 2011, pages 29–39.
- [123] W. Moncur and G. Leplâtre. Pictures at the ATM: Exploring the usability of multiple graphical passwords. In *Proceedings of the International Conference on Human Factors in Computing Systems (CHI'07)*. ACM, 2007, pages 887–894.
- [124] F. Monroe and M. K. Reiter. Graphical passwords. In L. F. Cranor and S. Garfinkel (editors). *Usability and Security*. O'Reilly, 2005, pages 147–164.
- [125] R. Morris and K. Thompson. Password security: A case history. *Communications of the ACM*, Vol. 22(11), ACM, 1979, pages 594–597.
- [126] H. Müller, J. Gove, and J. Webb. Understanding tablet use: A multi-method exploration. In *Proceedings of the International Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI '12)*. ACM, 2012, pages 1–10.
- [127] V. Müller. A survey on fallback authentication. *Bachelor Thesis at the University of Munich (LMU)*, 2014.
- [128] I. Muslukhov, Y. Boshmaf, C. Kuo, J. Lester, and K. Beznosov. Understanding users' requirements for data protection in smartphones. In *Proceedings of the International Conference on Data Engineering Workshops (ICDEW'12)*. IEEE, 2012, pages 228–235.
- [129] D. L. Nelson, V. S. Reed, and J. R. Walling. Pictorial superiority effect. *Journal of Experimental Psychology: Human Learning and Memory*, Vol. 2(5), American Psychological Association, 1976, pages 523–528.
- [130] A. Nosseir, R. Connor, and M. Dunlop. Internet authentication based on personal history – A feasibility test. In *Proceedings of Customer Focused Mobile Services Workshop at WWW'05*. ACM, 2005.
- [131] A. Nosseir, R. Connor, C. Revie, and S. Terzis. Question-based authentication using context data. In *Proceedings of the Nordic Conference on Human-computer Interaction (NordiCHI '06)*. ACM, 2006, pages 429–432.
- [132] L. O'Gorman, A. Bagga, and J. Bentley. Call center customer verification by query-directed passwords. *Financial Cryptography*, Vol. 3110, Springer, 2004, pages 54–67.

-
- [133] L. O'gorman, A. Bagga, and J. Bentley. Query-directed passwords. *Computers & Security*, Vol. 24(7), Elsevier, 2005, pages 546–560.
 - [134] M. Oka, K. Kato, Y. Xu, L. Liang, and F. Wen. Scribble-a-secret: Similarity-based password authentication using sketches. In *Proceedings of the International Conference on Pattern Recognition (ICPR'08)*. IEEE, 2008, pages 1–4.
 - [135] C. Park, J. Paik, T. Choi, S. Kim, Y. Kim, and J. Namkung. Multi-modal human verification using face and speech. In *Proceedings of the International Conference on Computer Vision Systems (ICVS'06)*. IEEE, 2006, pages 448–456.
 - [136] J. Podd, J. Bunnell, and R. Henderson. Cost-effective computer security: Cognitive and associative passwords. In *Proceedings of the Australian Conference on Computer-Human Interaction (OzCHI'96)*. IEEE, 1996, pages 304–305.
 - [137] I. Polakis, M. Lancini, G. Kontaxis, F. Maggi, S. Ioannidis, A. D. Keromytis, and S. Zanero. All your face are belong to us: Breaking facebook's social authentication. In *Proceedings of the Annual Conference on Computer Security Applications Conference (ACSAC'12)*. ACM, 2012, pages 399–408.
 - [138] R. Pond, J. Podd, J. Bunnell, and R. Henderson. Word association computer passwords: The effect of formulation techniques on recall and guessing rates. *Computers & Security*, Vol. 19(7), Elsevier, 2000, pages 645 – 656.
 - [139] A. Rabkin. Personal knowledge questions for fallback authentication: Security questions in the era of facebook. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS'08)*. ACM, 2008, pages 13–23.
 - [140] A. Ramirez, J. B. Walther, J. K. Burgoon, and M. Sunnafrank. Information-seeking strategies, uncertainty, and computer-mediated communication. *Human Communication Research*, Vol. 28(2), Blackwell Publishing Ltd, 2002, pages 213–228.
 - [141] M. Rasmussen and F. W. Rudmin. The coming PIN code epidemic: A survey study of memory of numeric security codes. *E-Journal of Applied Psychology*, Vol. 6(2), APA, 2010, pages 5–9.
 - [142] K. Renaud and A. De Angeli. My password is here! An investigation into visuo-spatial authentication mechanisms. *Interacting with Computers*, Vol. 16(6), Oxford University Press, 2004, pages 1017–1041.
 - [143] K. Renaud and A. De Angeli. Visual passwords: Cure-all or snake-oil? *Communications of the ACM*, Vol. 52(12), ACM, 2009, pages 135–140.
 - [144] K. Renaud and M. Just. Pictures or questions? examining user responses to association-based authentication. In *Proceedings of the British Computer Society Interaction Specialist Group Conference (BCS'10)*. British Computer Society, 2010, pages 98–107.

- [145] M. Richter. Fallback authentication based on location and life events. *Bachelor Thesis at the University of Munich (LMU)*, 2014.
- [146] B. L. Riddle, M. S. Miron, and J. A. Semo. Passwords in use in a university timesharing environment. *Computer & Security*, Vol. 8(7), Elsevier, 1989, pages 569–579.
- [147] J. Rokita, A. Krzyżak, and C. Y. Suen. Cell phones personal authentication systems using multimodal biometrics. In *Proceedings of the International Conference on Image Analysis and Recognition (ICIAR'08)*. Springer, 2008, pages 1013–1022.
- [148] B. Rysgaard. A method for protecting user data stored in memory of a mobile communication device, particularly a mobile phone, 2001. European Patent No. EP 1107627.
- [149] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Transactions on ComputersAcoustics, Speech and Signal Processing*, Vol. 26(1), IEEE, 1978, pages 43–49.
- [150] C. Saliger. Fallback authentication based on icon arrangement. *Bachelor Thesis at the University of Munich (LMU)*, 2014.
- [151] M. Satyanarayanan. Swiss army knife or wallet? *Pervasive Computing*, Vol. 4(2), IEEE, 2005, pages 2–3.
- [152] M. Sauter. *Grundkurs Mobile Kommunikationssysteme: UMTS, HSDPA, LTE, GSM, GPRS und Wireless LAN*. Vieweg & Sohn, 2006.
- [153] F. Schaub, R. Deyhle, and M. Weber. Password entry usability and shoulder surfing susceptibility on different smartphone platforms. In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia (MUM'12)*. ACM, 2012, pages 13:1–13:10.
- [154] S. Schechter, A. J. B. Brush, and S. Egelman. It's no secret: Measuring the security and reliability of authentication via 'secret' questions. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS'09)*. ACM, 2009, pages 40:1–40:1.
- [155] S. Schechter, S. Egelman, and R. W. Reeder. It's not what you know, but who you know: A social approach to last-resort authentication. In *Proceedings of the International Conference on Human Factors in Computing Systems (CHI'09)*. ACM, 2009, pages 1983–1992.
- [156] S. Schechter and R. W. Reeder. 1 + 1 = you: Measuring the comprehensibility of metaphors for configuring backup authentication. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS'09)*. ACM, 2009, pages 9:1–9:31.
- [157] J. Seifert, A. De Luca, and E. Rukzio. Don't queue up! user attitudes towards mobile interactions with public terminals. In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia (MUM'12)*. ACM, 2012, pages 45:1–45:4.

-
- [158] A. Shamir. How to share a secret. *Communications of the ACM*, Vol. 22(11), ACM, 1979, pages 612–613.
 - [159] 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 Symposium on Usable Privacy and Security (SOUPS'10)*. ACM, 2010, pages 2:1–2:20.
 - [160] E. Shi, Y. Niu, M. Jakobsson, and R. Chow. Implicit authentication through learning user behavior. In *Proceedings of the International Conference on Information Security (ICIS'11)*. Springer, 2011, pages 99–113.
 - [161] S. L. Smith. Authenticating users by word association. *Computers & Security*, Vol. 6(6), IEEE, 1987, pages 464 – 470.
 - [162] A. Sotirakopoulos. Motivating users to choose better passwords through peer pressure. *Poster at the Symposium on Usable Privacy and Security (SOUPS'11)*, 2011.
 - [163] A. Spink, B. Jansen, D. Wolfram, and T. Saracevic. From E-sex to E-commerce: Web search changes. *Computer*, Vol. 35(3), IEEE, 2002, pages 107–109.
 - [164] J. Spitzer, C. Singh, and D. Schweitzer. A security class project in graphical passwords. *Journal of Computing Sciences in Colleges*, Vol. 26(2), Consortium for Computing Sciences in Colleges, 2010, pages 7–13.
 - [165] L. R. Squire. Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. *Journal of Cognitive Neuroscience*, Vol. 4(3), MIT Press, 1992, pages 232–243.
 - [166] E. Stobert and R. Biddle. The password life cycle: User behaviour in managing passwords. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS'14)*. USENIX Association, 2014, pages 243–255.
 - [167] H. Sun, K. Wang, X. Li, N. Qin, and Z. Chen. Passapp: My app is my password! In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '15)*. ACM, 2015, pages 306–315.
 - [168] H.-M. Sun, Y.-H. Chen, C.-C. Fang, and S.-Y. Chang. Passmap: A map based graphical-password authentication system. In *Proceedings of the Symposium on Information, Computer and Communications Security (ASIACCS '12)*. ACM, 2012, pages 99–100.
 - [169] X. Suo, Y. Zhu, and G. S. Owen. Graphical passwords: A survey. In *Proceedings of the Annual Computer Security Applications Conference (ACSAC'05)*. IEEE, 2005, pages 463–472.
 - [170] Symantec. The symantec smartphone honey stick project report, 2011.

- [171] A. F. Syukri, E. Okamoto, and M. Mambo. A user identification system using signature written with mouse. *Information Security and Privacy*, Vol. 1438, Springer, 1998, pages 403–414.
- [172] M. Tamviruzzaman, S. I. Ahamed, C. S. Hasan, and C. O'brien. ePet: When cellular phone learns to recognize its owner. In *Proceedings of the Workshop on Assurable and Usable Security Configuration (SafeConfig '09)*. ACM, 2009, pages 13–18.
- [173] S. A. Thalhammer. A context-aware backup authentication system for smartphones. *Bachelor Thesis at the University of Munich (LMU)*, 2013.
- [174] J. Thorpe, B. MacRae, and A. Salehi-Abari. Usability and security evaluation of geopass: A geographic location-password scheme. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS'13)*. ACM, 2013, pages 14:1–14:14.
- [175] E. Tulving. Episodic and semantic memory. In E. Tulving and W. Donaldson (editors). *Organization of Memory*. Academic Press, 1972, pages 381–403.
- [176] E. Tulving. What is episodic memory? *Current Directions in Psychological Science*, Vol. 2(3), Sage Publications, 1993, pages 67–70.
- [177] 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 *Proceedings of the Conference on Security Symposium (SS'12)*. USENIX Association, 2012, pages 65–80.
- [178] 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 Symposium on Usable Privacy and Security (SOUPS'13)*. ACM, 2013, pages 10:1–10:14.
- [179] C. Varenhorst, M. Kleek, and L. Rudolph. Passdoodles: A lightweight authentication method. *Research Science Institute*, 2004.
- [180] E. von Zezschwitz, A. De Luca, and H. Hussmann. Survival of the shortest: A retrospective analysis of influencing factors on password composition. In *Proceedings of the International Conference on Human-Computer Interaction (Interact'13)*. Springer, 2013, pages 460–467.
- [181] 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 Nordic Conference on Human-Computer Interaction (NordiCHI'14)*. ACM, 2014, pages 461–470.
- [182] K.-P. L. Vu, R. W. Proctor, A. Bhargav-Spantzel, B.-L. B. Tai, J. Cook, and E. E. Schultz. Improving password security and memorability to protect personal and organizational information. *International Journal of Human-Computer Studies*, Vol. 65(8), Academic Press, 2007, pages 744 – 757.

-
- [183] 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 Conference on Computer and Communications Security (CCS'10)*. ACM, 2010, pages 162–175.
 - [184] M. Weir, S. Aggarwal, B. de Medeiros, and B. Glodek. Password cracking using probabilistic context-free grammars. In *Proceedings of the Symposium on Security and Privacy (SP '09)*. IEEE, 2009, pages 391–405.
 - [185] R. Weiss and A. De Luca. PassShapes: Utilizing stroke based authentication to increase password memorability. In *Proceedings of the Nordic Conference on Human-computer Interaction (NordiCHI'08)*. ACM, 2008, pages 383–392.
 - [186] S. Wiedenbeck, J. Waters, J.-C. Birget, A. Brodskiy, and N. Memon. PassPoints: Design and longitudinal evaluation of a graphical password system. *International Journal of Human-Computer Studies*, Vol. 63(1), Academic Press, 2005, pages 102–127.
 - [187] H. M. Wood. The use of passwords for controlling access to remote computer systems and services. In *Proceedings of the American Federation of Information Processing Societies*. ACM, 1977, pages 27–33.
 - [188] J. Yan, A. Blackwell, R. Anderson, and A. Grant. Password memorability and security: Empirical results. *Security & Privacy*, Vol. 2(5), IEEE, 2004, pages 25–31.
 - [189] S. Yardi, N. Feamster, and A. Bruckman. Photo-based authentication using social networks. In *Proceedings of the First Workshop on Online Social Networks (WOSN'08)*. ACM, 2008, pages 55–60.
 - [190] M. Zviran and W. J. Haga. Cognitive passwords: The key to easy access control. *Computers & Security*, Vol. 9(8), Elsevier, 1990, pages 723 – 736.
 - [191] M. Zviran and W. J. Haga. A comparison of password techniques for multilevel authentication mechanisms. *The Computer Journal*, Vol. 36(3), Oxford University Press, 1993, pages 227–237.

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 09. Dezember 2015

Alina Hang